Statistical Models and Methods for Dependent Life History Processes

by

Jooyoung Lee

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics - Biostatistics

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Susan Murray Professor, Dept. of Biostatistics, University of Michigan
Supervisor(s):	Richard J. Cook Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal Member:	Jerald Lawless Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal Member:	Leilei Zeng Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Ashok Chaurasia Professor, School of Public Health and Health Systems, University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis deals with statistical issues in the analysis of complex life history processes which have characteristics of heterogeneity and dependence. We are motivated, in this thesis, by three specific types of processes; i) processes featuring recurrent episodic conditions ii) multi-type recurrent events, and iii) clustered multi state processes as arise in family studies.

In chronic diseases featuring recurrent episodic conditions, symptom onset is followed by a period during which symptoms are present until recovery. In the analysis of data from such processes, analysis is often based only on the recurrent onset of disease, ignoring the duration of symptoms. This loss of information may lead to incorrect conclusions in the analysis of this data. In Chapter 2, we propose a novel model for an alternating two-state process including symptom-free state and symptomatic state to recognize the duration of symptoms. This approach reflects the dynamics of individual's disease process and helps to understand a course of disease. Intensity-based models with multiplicative random effects are considered where the disease onset time is governed by a conditionally Markov intensity and the time of recovery is governed by a conditionally semi-Markov intensity. A bivariate random effect with one multiplicative component for each intensity is introduced to accommodate between-individual heterogeneity and a dependence between bivariate random effect variables offers a natural and more general framework for modeling the two state process. A copula function is used for the joint distribution of random effects which retains the marginal features and gives flexible choices of dependence structure. The proposed model is a semiparametric model for which estimation is carried out using an expectation-maximization algorithm.

The aforementioned problem leads us to investigate the impact of ignoring symptom duration in a randomized trial setting. In Chapter 3, we define two risk sets for recurrent event analyses: one involves including individuals during their symptomatic period, and the other excluding individuals from the risk set during symptomatic periods. In a clinical trial, the balance between treatment groups in unmeasured confounders present at the time of randomization can be lost following randomization as the risk set changes, thus, retaining individuals in the risk set is a common approach. Here we examine asymptotic

and empirical biases of estimators from the rate-based models when two different risk sets are applied. We assume that the true underlying process is an alternating two-state process where the true risk set is the one that excludes individuals when they are experiencing an exacerbation. We consider two scenarios of the true model. First, there is no between-variation for each process and no dependence between two processes. The second scenario is to use the proposed dependent alternating two-states model in Chapter 2. Issues of model misspecification and causal inference are considered. When focus is on clinical trials, power implications of risk set misspecification is of interest.

In Chapter 4, attention is directed at multiple recurrent events where each endpoint is of interest. The use of composite endpoint which is the time point of the first event of any type is a simple way to analyse such data. However, when multiple events are of comparable importance, use of a composite endpoint analysis may not be suitable. We propose a copula-based model for multi-type recurrent events where each type of recurrent event process arises from a mixed-Poisson model and random effects linking the events through a copula function. When more than two types of events are considered, composite likelihood is adopted to ease the computational burden, and simultaneous and two-stage estimation are explored.

An aim of family studies is typically to gain knowledge about factors governing the inheritance of diseases. One may be interested in examining a dependence of disease onset between family members, and in identifying genetic markers associated with heritable disease. A common procedure is to collect families is through probands in which such affected individuals are selected from a disease registry and their family members (non-probands) are, then, recruited for examination. This approach to sampling families motivates us to consider the disease onset process along with survival since the proband must be diseased and alive to be recruited, and family members may need to be alive. In Chapter 5, we propose a model for a clustered illness-death process for family studies which accounts for the semi-competing risks problem for disease onset as well as biased sampling. We model within-family association in the age of disease onset via a copula function and applied to the possibly latent disease onset time and incorporate survival through a marginal illness-death model. The ascertainment condition is reflected in the likelihood or composite likelihood construction. Two study designs regarding the recruitment of family members are consid-

ered. One involves the collection of disease history from family members via the proband or medical records. The other requires family members to undergo a medical examination in which case they must be alive at the time of the family study. Family data alone are insufficient to estimate all of the parameters of the illness-death processes. We therefore make use of auxiliary data including the population mortality data and additional registry data to address the estimatability issue. Another source of auxiliary data is current status survey. The issue of missing genetic markers is also addressed in each study design.

Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Dr. Richard J. Cook, for his exceptional guidance, support, and encouragement. He has guided me through the field of Biostatistics and has taught me extensively with patience. This work would not have been possible without his enormous care and mentorship.

I would like to thank my thesis committee, Dr. Jerry Lawless, Dr. Leilei Zeng, Dr. Ashok Chaurasia, and Dr. Susan Murray for their valuable comments and feedbacks.

A very special gratitude goes out to Ms. Mary Lou Dufton and Ms. Marg Feeney for their help in every aspect. I also wish to thank Ker-Ai Lee for her advice with computing.

And finally, last but by no means least, I am very grateful and thankful for my parents, my brother, and my dearest and best friend, Justin S. Lee, for their endless support and love during my doctoral study.

Dedication

To my family Eungjin Lee, Hyesuk Jung, and Seunghyun Lee

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Individuals experience events during their lifetimes, and it is important to analyze such data to understand the processes governing events occurrence. The methods for analysis depend on the nature of data and how that data are acquired and such issues are particularly important for the analysis of life history data. In many settings, it is of interest to study the dynamics of disease processes over the course of an individual's lifetime, variation in patterns across individuals, how interventions may affect such processes, and relationships between more than two processes. From a statistical point of view, estimation of the probability of disease incidence and event occurrence, covariate effects, and measures of dependence are often of interest. This thesis is concerned with three different problems: i) analysis of recurrent episodic conditions reflecting the onset and duration of symptomatic periods in studies of chronic diseases, ii) multiple recurrent events possibly arising due to the same underlying cause which are therefore associated, and iii) clustered multi-state data arising in the conduct of family studies. Dependence modeling and dealing with heterogeneity are themes in each of these problems. Other themes are the impact of biased sampling schemes and the use of auxiliary data which arise in the third problem.

This introductory chapter begins with an overview of statistical methods for lifetime

data, describes the motivating problems which are used in the thesis, and briefly discuss the contents of the thesis.

1.2 Brief Overview of Lifetime Data Analysis

Methods for the analysis of lifetime processes have been extensively developed to deal with problems in medicine, economics, actuarial science, and engineering. Examples of such processes include in health research the onset and progression of cancer, recurrent hospitalizations, destruction of joints over time in arthritis, etc. In this thesis, we focus on recurrent event processes and multistate processes. There are several frameworks for recurrent event data analysis including intensity-based methods, rate-based models and random effects models (Lawless and Nadeau, 1995; Cook and Lawless, 2007). We review these briefly in Section 1.2.1. Multi-state processes are of use for competing risks or the study of progressive or reversible disease processes. We review related statistical methods for multi-state processes in Section 1.2.2.

1.2.1 Recurrent Event Data

Recurrent event processes generate events repeatedly over time. Such processes arise in many fields within health research including the occurrence of asthma attacks in respirology trials, epileptic seizures in neurology studies, and recurrent hospitalizations in affective disorders (Cook and Lawless, 2007). Statistical models for recurrent events can be specified in terms of intensity functions for point processes which offers a useful framework for the analysis of such data (Andersen *and others*, 1993).

Let T_{ik} denote the time of the kth event for individual i in a sample of n independent individuals, $i = 1, \ldots, n$. Let $N_i(t) = \sum_{k=1}^{\infty} I(T_{ik} \leq t)$ count the number of events occurring over $[0, t]$ for individual i where $dN_i(t) = 1$ if an event occurs at time t, and $dN_i(t) = 0$ otherwise. If $X_i(t)$ is a vector of covariates for individual i, $\{X_i(t), 0 < t\}$ denotes the covariate process. We let $H_i(t) = \{N_i(s), X_i(s), 0 \leq s < t\}$. If we let C_i denote the right censoring time, $Y_i(t) = I(t \leq C_i)$ is an at risk indicator for the event. Throughout this

thesis we suppose the observation process $\{Y_i(t), 0 \leq t\}$ is independent of the event process $\{dN_i(t), 0 \leq t\}$ given $X_i(s)$. We let $d\bar{N}_i(t) = Y_i(t)dN_i(t)$ and $\bar{N}_i(t) = \int_0^t d\bar{N}_i(s)$.

Intensity Functions

Let $\bar{H}_i(t) = \{Y_i(s), \bar{N}_i(s), X_i(s), 0 \leq s < t\}$ denote the history of the observation, observed event process, and external covariate process at time t; we assume for simplicity the covariate process is external. Then the intensity function is defined as the instantaneous probability of event occurrence at time t given the process history, and is written as

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_i(t) = 1 | \bar{H}_i(t))}{\Delta t} = Y_i(t)\lambda(t|H_i(t))$$

where $\Delta \bar{N}_i(t) = \bar{N}_i(t + \Delta t^-) - \bar{N}_i(t^-)$ is the observed number of events over the interval $[t, t + \Delta t)$. Here

$$\lambda(t|H_i(t)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_i(t) = 1 | H_i(t))}{\Delta t}$$

is the intensity function for the underlying recurrent event process. Full specification of the intensity function requires one to condition on the history of event process. For the examination of treatment effects in clinical trials this therefore is not an ideal basis for analysis (Kalbfleisch and Prentice, 2011).

Under the assumption of modulated Poisson processes, the rate function given fixed covariates, say, where $H_i(t) = \{N_i(s), 0 \leq s < t, X_i\}$ is defined as

$$\lambda(t|H_i(t)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_i(t) = 1 | H_i(t))}{\Delta t} = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_i(t) = 1 | X_i)}{\Delta t} = \rho_i(t),$$

and the corresponding mean function is $\mu_i(t) = E[N_i(t)|X_i] = \int_0^t \rho_i(s)ds$. The most common framework for modeling covariate effects is multiplicative models of the form

$$\rho_i(t) = \rho_0(t) \exp(x_i'\beta).$$

The baseline rate function can be parametrized or semiparametric models can be specified. Andersen and Gill (1982) proposed a semiparametric Cox-type model and provided the

large sample properties of estimators through martingale theory. The partial likelihood score function is defined as

$$U(\beta, t) = \sum_{i=1}^{n} \int_0^t Y_i(u) \{x_i(u) - \bar{x}(\beta, u)\} \, dN_i(u) \tag{1.1}$$

where $\bar{x}(\beta, t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$ with $S^{(k)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) x_i(t)^{\otimes k} \exp(x_i'(t)\beta), k = 0, 1, 2$.

Random Effects Models

To accommodate between-subject variability, we introduce random effects into the Poisson model to obtain a mixed Poisson model of the form

$$\lambda(t|H_i(t), u_i) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_i(t) = 1|H_i(t), u_i)}{\Delta t} = u_i \rho_i(t)$$

where U_i is an unobservable nonnegative random variable independent of $X_i(t)$ with mean 1 and variance ϕ. Note that the interpretation of covariates effects are conditional on random effects u_i, which means that the marginal covariate effects are more complicated. The unconditional intensity function for a mixed Poisson process is of the form

$$\lambda(t|H_i(t)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_i(t) = 1|H_i(t))}{\Delta t} = \rho_i(t) E[U_i|H_i(t)]$$

which depends on the precise random effects distribution adopted. A number of distributions for U_i are available with commonly adopted ones including the gamma, inverse Gaussian, positive stable, and log-normal distributions (Hougaard, 2000).

Robust Marginal Methods

If interest lies in marginal features such as the rate or mean functions in order to assess covariate effects, robust methods are often preferable (Cook and Lawless, 2007). Lin *and others* (2000) proposed semiparametric multiplicative models for the mean and rate

functions for the counting process with fixed covariates $x_i(t) = x_i$. This has the form

$$E[dN_i(t)|Y_i(t), x_i)) = \rho_0(t)dt \exp(x_i'\beta).$$

The estimating equation for β is equivalent to (1.1) under the Poisson assumption. To provide protection against extra-Poisson variation or other forms of misspecification, robust variance estimates are required for valid inferences.

1.2.2 Multi-state Processes

A multi-state model is a model for a stochastic process with a discrete states which is often used to describe life history processes changing over a period of time (Hougaard, 1999). It is a useful framework describing the progression of retinopathy in diabetes, joint damage in psoriatic arthritis, or cancer and death in oncology (Cook and Lawless, 2018). Multi-state processes often enable one to calculate transition probabilities over a period of time, or marginal state occupancy probabilities. We introduce various multi-state models in this section. Note that we are primarily concerned with fixed covariates here.

Intensity functions

Let $Z_i(t)$ denote the state occupied at time t for an individual i with state space $\{1, 2, \ldots, K\}$. and $H_i(t) = \{Z_i(u), 0 \leq u < t, x_i\}$ denote the history of state occupancy over $[0, t]$ where x_i is a vector of fixed covariates. Then, the intensity function of state k to l transitions is defined as

$$\lim_{\Delta t \downarrow 0} \frac{P(Z_i(t + \Delta t^-) = l | Z_i(t^-) = k, H_i(t))}{\Delta t} = \lambda_{kl}(t|H_i(t))$$

for all $k \neq l$ and $k, l \in \{1, \ldots, K\}$. We can construct the likelihood for a specific individual sample path using intensity functions.

Markov and Semi-Markov Models

Markov processes are commonly used, for which the transition intensities depend on only the state currently occupied. In this setting, the time scale is the global time (*i.e.* the time since the origin of the process). Then, the transition intensities are of the form

$$\lambda_{kl}(t|H_i(t)) = Y_{ik}(t)\lambda_{kl}(t|x_i)$$

where $Y_{ik}(t) = I(Z_i(t^-) = k), k = 1, \ldots, K$. If we let $\Lambda_{kl}(t|x_i) = \int_0^t \lambda_{kl}(u|x_i)du$, then the transition probability matrix $P(s,t|x_i)$ can be obtained by product integration as

$$P(s,t|x_i) = \prod_{(s,t]} \{\mathcal{I} + d\Lambda(u|x_i)\}$$

where $\Lambda(t|x_i)$ is a $K \times K$ matrix with $\Lambda_{kl}(t|x_i)$ in the(k,l) entry, $j \neq k$, $-\sum_{l \neq k} \Lambda_{kl}(t|x_i)$ in the diagonal entries, and $\mathcal{I}$ is an identity matrix of size K (Andersen *and others*, 1993).

Sometime it is more natural to use the time since the most recent transition as the time scale. Examples of such settings include studies of the duration for recurrent infections in chronic bronchitis, studies of the duration of recurrent hospitalizations, and the duration of depressive episodes among individuals with affective disorder. For semi-Markov processes,

$$\lambda_{kl}(t|H_i(t)) = Y_{ik}(t)h_{kl}(B_{ik}(t)|x_i),$$

where $B_{ik}(t)$ is the time since entry to the current state k.

Heterogeneity and Dependence Modeling

Sometimes individual life history paths exhibit substantial heterogeneity defined by the presence of considerably more variation than can be accounted for by some base model. In this case random effects are often useful to accommodate between-subject variation. This approach is often used in life history as shown in Section 1.2.1 with the conditional intensity given individual unobservable variables. In a similar fashion with multiplicative

models, the conditional transition intensities are of the form

$$\lambda_{kl}(t|H_i(t), u_{ikl}) = u_{ikl}Y_{ik}(t)\lambda_{kl}(t|x_i)$$

where U_{ikl} is a latent non-negative random variable introduced to account for variation between individuals, and u_{ikl} is its realized value.

When an individual experiences multiple processes, interest may lie in jointly modeling them and measuring the dependence between them. One approach readily adopted involves random effects where the multiple processes are linked via latent variables. While quite common, this framework does not provide easily interpretable marginal covariate effects or measures of association. Copula models offer a natural alternative where marginal processes are jointly modeled by copula functions (Joe, 1997). This formulation determines a dependence structure only through the copula while retaining the simplicity of the marginal models.

A copula function $C(v_1, \ldots, v_K)$ in K dimensions is defined as a multivariate distribution function with marginal uniform $[0, 1]$ distribution; *i.e.* $V_1, \ldots, V_K$ are uniform $[0, 1]$ random variables (Joe, 1997). We can then write

$$\mathcal{C}(v_1, \ldots, v_K) = P(V_1 \leq v_1, \ldots, V_K \leq v_K).$$

For simplicity of illustration, we consider the progressive $K + 1$ state process and let W_k be a the sojourn time in state k, $k = 1, \ldots, K$ (He and Lawless, 2003). Let $\mathcal{F}_k(t|x)$ be the survival function of W_k given $X = x$. Then the joint survival function $\mathcal{F}(w_1, \ldots, w_K|x) = P(W_1 \geq w_1, \ldots, W_K \geq w_K|x)$ can be specified via a copula as

$$\mathcal{F}(w_1, \ldots, w_K|x) = \mathcal{C}(\mathcal{F}_1(w_1|x), \ldots, \mathcal{F}_K(w_K|x)).$$

As a commonly used measure of dependence for two pairs of random variables (T_{1j}, T_{1k}) and (T_{2j}, T_{2k}), Kendall's τ is defined as

$$\tau = P((T_{1j} - T_{2j})(T_{1k} - T_{2k}) > 0) - P((T_{1j} - T_{2j})(T_{1k} - T_{2k}) < 0),$$

and this is often used in copula modeling as the summary dependence measure. One of the widely used classes of copula function is the Gaussian copula, which has the form

$$\mathcal{C}(v_1, \ldots, v_K) = \Phi_R(\Phi^{-1}(v_1), \ldots, \Phi^{-1}(v_K)),$$

where Φ_R is a joint cumulative distribution function of a multivariate normal distribution with mean zero and correlation matrix R and $\Phi(\cdot)$ is the standard normal distribution function. The Archimedean copula family is also popular, in which the copula functions have the form

$$\mathcal{C}(v_1, \ldots, v_K) = \psi^{-1}(\psi(v_1; \phi) + \ldots + \psi(v_K; \phi); \phi),$$

where $\psi(\cdot; \phi)$ is a so-called the generator function and ϕ is a dependence parameter. The Clayton copula is a member of the Archimedean family (Genest and Rivest, 1993), given by

$$C(v_1, \ldots, v_K) = (v_1^{-\phi} + \ldots + v_K^{-\phi} - K + 1)^{-1/\phi}.$$

1.3 Studies and Motivating Applications

Here we will briefly describe some motivating problems which will be revisited in this thesis.

1.3.1 Recurrent Hospitalizations in Affective Disorder

The data of a registry of recurrent hospitalizations in Denmark were collected over the period 1994-1999 to study affective disorder (Kessing *and others*, 2004). All patients who entered the study had been diagnosed with having affective disorder and being hospitalized at least once between 1994 and 1999. This selection condition was applied to a total of 10,523 patients. The number of males is 6,721 (63.9%) and the number of females is 3,802 (36.1%). Over the study window the average number of admissions is 1.62 (SD = 1.72), with a minimum of 1 and a maximum of 90. The data is collected prospectively following the start of the first hospitalization.

Figure 1.1 illustrates profiles for a sample of 10 patients. The solid line indicates the

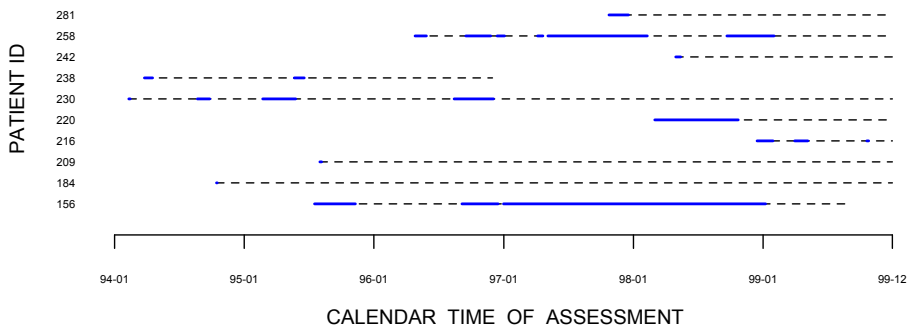


Figure 1.1: Profiles of patients from the Danish psychiatric hospitalization study

state of being in hospital and the dotted line indicates state of being out of hospital. From this plot, we observe that some hospitalizations are of an appreciable duration, which motivates the use of an alternating two-state process. The data were censored at the time of the end of the study on December 31, 1999, at the time of death, or if patients were diagnosed with organic disorder, schizophrenia or schizo-affective disorder. The classification of the type of affective disorder was made at the date of discharge from the hospital, thus the type of disorder was time-dependent; patients were hospitalized with a diagnosis of unipolar or a diagnosis of bipolar affective disorder. A total of 9417 patients had been diagnosed with unipolar disorder and 1106 patients had been diagnosed with mania or bipolar disorder at the first discharge. The number of total re-admissions is 6498 and the the duration of the subsequent hospitalizations varied from 0 to 1253 days (mean = 43.9, SD =56.4) (Kessing *and others*, 2004). We revisit this study in Chapter 2 to illustrate the fitting of an alternating two-state model we develop in that Chapter.

1.3.2 A Herpes Simplex Trial

Herpes simplex virus infection causes recurrent outbreaks of symptoms lasting typically two to four weeks of duration. A multicenter open-label randomized two-period crossover trial was conducted to compare the efficacy of suppressive therapy versus episodic therapy (Romanowski *and others*, 2003). Suppressive therapy was valacyclovir at a dosage of 500 mg once daily and episodic therapy was valacyclovir at a dosage of 500 mg twice daily for 5 days after the outbreaks of symptoms. If herpes outbreaks in the suppressive arm patients received episodic therapy for 5 days and returned to suppressive therapy after 5 days. A total of 202 patients completed the study for two 24-week period of study out of a total of 225 patients at enrollment. After the first period of study patients switched to another therapy so that each patients received both treatments for the 48-week study period. The mean of total number of outbreaks for the first period is 4.019 with the standard error of 3.898. The mean of symptom duration is 24.1 days with a minimum of 1 day and a maximum of 175 days. The variation of the duration of symptom inspires us to investigate different approaches to defining the risk sets (i.e. including or excluding individuals when they are experiencing events). We revisit this study in Chapter 3.

1.3.3 Iron Supplementation in Malnourished Children

Malnutrition in children in low-income countries has been identified as a cause of immune deficiency and susceptibility to infectious diseases since activation of the immune system in response to infection requires additional energy. Examples of infectious diseases arising due to malnutrition are opportunistic pathogens and fungus, noma, respiratory, intestinal infections, tuberculosis, measles and other chronic infections (Ambrus, 2004; Schaible and Stefan, 2007). Iron deficiency, which is also prevalent in developing countries, causes anemia and a deficiency of red blood cells. Lemaire *and others* (2011) conducted a study to examine the efficacy of iron-containing micro-nutrient powder (iron MNP) on the risk of infections in malnourished children. In a randomized clinical trial, 268 Bangladeshi children, aged 12-24 month, and moderately-to-severely malnourished with a hemoglobin concentration between 70 and 110 g/L, were recruited in two phases, 12/2007-06/2008 and 07/2008-01/2009, respectively. Iron MNP were provided to 136 children daily for 2 months and the remaining

children were provided placebo powder. The primary endpoint was the occurrence of infections and associated symptoms such as diarrhea, dysentery, lower respiratory tract infections (LRTIs), cough and fever. During the 2 months intervention period, the incidence of infections was assessed every other day, whereas after the intervention period, it was assessed weekly. Interest may lie in measuring dependence between the onset of multiple infections, which motivates a joint model for multi-type recurrent events. We revisit this study in Chapter 4 to illustrate the use of a multi-type recurrent event model.

1.3.4 Psoriatic Arthritis Family Study

PsA(psoriatic arthrists) is an immune-mediated inflammatory disease occurring commonly in patients with psoriasis. Its symptoms include peripheral joints and spinal pain or stiffness, enthesis, and dactylitis (Gladman, 1991). The Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto maintains a Psoriatic Arthritis Clinic, which was established in 1976 and has been following patients since its formation. Upon entry to the clinic, patients undergo a detailed examination and provide serum samples. Follow-up clinical and radiological assessments are scheduled annually and biannually to track the changes in joint damage and functional ability, and serum samples are taken at clinic visits to measure the changes of markers (Cook and Lawless, 2014). As of April of 2017, 1436 patients have been recruited to University of Toronto Psoriatic Arthritis Registry (UTPAR) with the range of the date of birth from 1893 to 1997. A median of age at the first assessment is 44.1 and the mean of age at PsA is 38.0 (SD = 13.6). A family study was carried out at this registry to examine familial aggregation in the occurrence of PsA. Among the 1436 individuals, 150 were selected for family studies as probands. Figure 1.2 gives an example Lexis diagram for a family in the PsA family studies. The proband was sampled from the UTPAR in 2001 and two parents of the proband were recruited to the family study in 2007; the father and mother were born in 1929 and 1934, respectively, and the proband (son) was born in 1955. We revisit this family study data in Chapter 5 where we develop methods for clustered family data where each individual follows a marginal illness-death process.

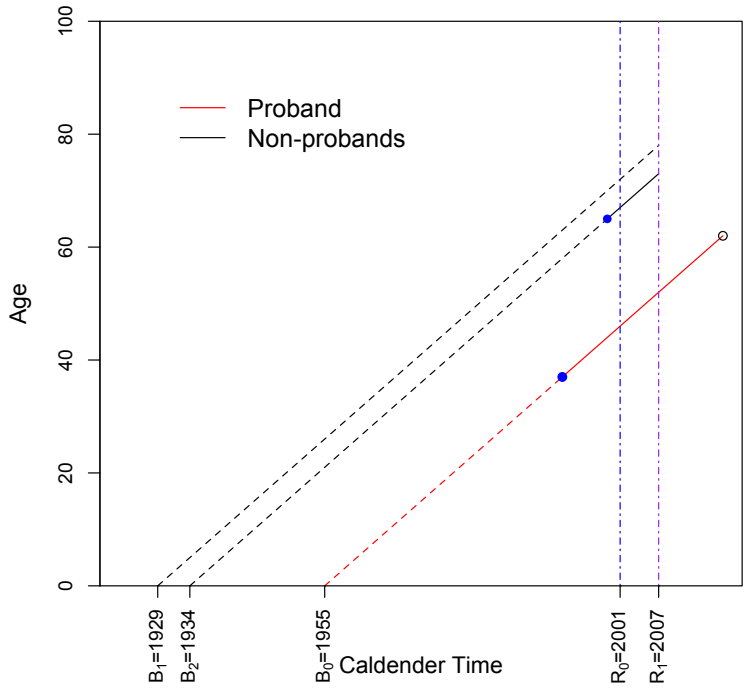


Figure 1.2: Lexis diagram for a family with 3 members; one proband and parents

1.4 Contents of the Thesis

In Chapter 2, we develop models for the analysis of alternating two-state processes, motivated by studies of chronic diseases in which affected individuals experience recurrent symptomatic periods, each of which may last for an appreciable time. We formulate a copula-based model to link a subject-specific multiplicative random effect acting on conditionally Markov intensity for the onset of exacerbations, with a random effect acting on a conditionally semi-Markov intensity for exacerbation durations. An expectation-maximization algorithm is described for fitting a joint semiparametric model for the onset and resolution of exacerbations. An application is given to a study of recurrent hospitalization in patients with affective disorder (Kessing *and others*, 2004).

While marginal rate-based analyses have considerable appeal for the analysis of recur-

rent exacerbation in clinical trials, relatively little work has been carried out on how to best handle the duration of symptomatic periods. In chapter 3, we derive the asymptotic bias of the Nelson-Aalen estimator of cumulative mean function as well as regression coefficients under an Andersen-Gill model as a function of parameters of an underlying two-state process. We investigate the impact of the mean duration of the exacerbation durations, heterogeneity, and dependence on the asymptotic and empirical biases of parameter estimates. An application to a trial of individuals with herpes simplex virus is given for illustration (Romanowski *and others*, 2003).

When several types of recurrent events may arise, interest often lies in marginal modeling and studying the nature of the dependence structure. In Chapter 4, we propose multivariate mixed-Poisson model with the dependence between event type-specific random effects accommodated through a Gaussian copula. Such models retain the simple interpretation of marginal features, separately reflect the heterogeneity in risk for each type of event, and provide insight into the dependence between the different types of events. Inference is proposed based on composite likelihood to avoid high dimensional integration. The relative efficiency of estimators obtained from simultaneous and two-stage estimation is examined. An application to a study of nutritional supplements in malnourished children is given in which the goal is to evaluate the reduction in the rate of several types of infection (Lemaire *and others*, 2011).

Chapter 5 considers family studies where accommodating within-family association in the age of disease onset is critical when studying the genetic basis of chronic disease. In family studies families are typically recruited by the identification of an affected individual from a disease registry, called the proband, whose disease onset time is right-truncated. The disease status, and possibly onset times of other family members, called non-probands, is then collected; sometimes this is retrospectively reported and sometimes non-probands are required to undergo examination to determine their disease status. Relatively little work has been done on the effect of mortality on inferences about the dependence of disease processes. We construct likelihoods for family data based on a marginal illness-death model where the joint distribution of the age at disease onset is considered under complex sampling schemes. When the disease is rare and data are insufficient, auxiliary data can be used to augment the likelihood and facilitate estimation. We apply the proposed methods

to the analysis of a family study of psoriatic arthritis carried out at the University of Toronto (Pollock *and others*, 2015; Zhong and Cook, 2016).

Chapter 6 reviews the conclusions of the thesis and discusses further research topics for each area.

Chapter 2

Heterogeneity and Dependence Modeling for Alternating Two-state Processes Via Copulas

2.1 Introduction

Many chronic diseases are characterized by recurrent periods of time during which symptoms are manifest. Examples include recurrent exacerbations in individuals with respiratory disease (Grossman *and others*, 1998), recurrent bouts of depression among individuals with affective disorder (Garber *and others*, 1988), or recurrent hospitalization in patients with cardiovascular disease (Borer *and others*, 2012). In such settings, maximum likelihood estimation requires joint modeling of both the onset and duration of recurrent symptomatic periods. For convenience we use the term "exacerbation" to represent the condition in which the disease is in an "active" state to represent infections, flares of symptoms, hospitalizations, etc.

Intensity functions play a central role in the analysis of data from multistate processes (Andersen *and others*, 1993) and intensity-based models can provide useful insight into factors governing event occurrence (Aalen *and others*, 2008). For processes featuring recurrent

alternating sojourns in states, however, several aspects must be addressed including the time scale of the intensity functions. For conditions such as chronic bronchitis or chronic obstructive pulmonary disease, there is a gradual increase in lung damage due to repeated exacerbations which increase the risk of symptom outbreaks over time. Likewise, for patients with cardiovascular disease, as the condition deteriorates individuals are at increased risk of hospitalization. We therefore consider a model in which the risk of symptom onset for each individual can be characterized by a modulated Markov intensity function to allow the risk to change as the time with the disease condition increases. When symptomatic periods arise, they often follow a natural course, as is the case with episodic infections leading to symptom exacerbation in respiratory disease. Moreover, upon the onset of an exacerbation standard interventions may be delivered to ameliorate symptoms and resolve the symptomatic period. In both cases the resolution process begins upon the onset of the symptomatic period motivating the use of a modulated semi-Markov model for the duration of symptomatic periods.

A second complication is that there can be considerable unexplained heterogeneity between individuals in the propensity for, and duration of, symptomatic periods. Moreover, individuals at higher risk of exacerbations may also tend to have shorter sojourn times in the exacerbation state; in studies of infectious disease this can arise if some individuals live in an area putting them at high risk of reinfection, for example. To accommodate this type of heterogeneity we introduce a bivariate random effect in which one component acts multiplicatively on the Markov intensity for the onset of symptoms and the other acts multiplicatively on the semi-Markov intensity for symptom duration.

The primary purpose of this Chapter is to present a model for an alternating two-state process which has a suitable time scale for the two intensities, accommodates heterogeneity in the propensity for and duration of symptomatic periods, and allows for correlated random effects for the two conditional intensities. The remainder of this chapter is organized as follows. In Section 2.2 the formulation of the model is given in detail including the copula model used to accommodate dependence in the random effects. The marginal likelihood is derived and an expectation-maximization algorithm (Dempster *and others*, 1977) is given to facilitate semi-parametric analysis; variance estimation is also given with details provided in the Appendix 2.A. Simulation studies are described and reported on

in Section 2.3 where we study the impact of misspecifying the random effect distribution. In Section 2.4 we fit the model to data from a Danish study on repeated hospitalizations in individuals with affective disorder (Kessing *and others*, 2004). Concluding remarks are given in Section 2.5.

2.2 Modeling Heterogeneous Hybrid Markov/Semi-Markov Processes via Copulas

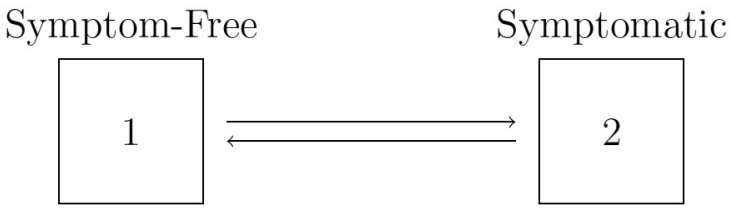


Figure 2.1: A two-state diagram for chronic diseases with recurrent symptomatic episodes

Consider the two-state diagram in Figure 2.1. We suppose each individual i in a sample of m independent individuals starts their process at time $t = 0$, which corresponds to the time of disease onset $i = 1, \ldots, m$. Let S_{ik} be the onset (start) time for the kth exacerbation, T_{ik} denote the resolution time of the kth exacerbation, and $W_{ik} = T_{ik} - S_{ik}$ the duration of kth exacerbation. The counting process $\{N_{ij}(u), 0 < u\}$ records the cumulative number of $j \to 3 - j$ transitions experienced by individual i over $(0, t]$, for $j = 1, 2$, where $N_{i1}(t) = \sum_{k=1}^{\infty} I(S_{ik} \leq t)$ and $N_{i2}(t) = \sum_{k=1}^{\infty} I(T_{ik} \leq t)$. Let $Z_i(s) = 1$ if individual i is symptom-free at $s > 0$, $Z_i(s) = 2$ if they are symptomatic, and $Y_{ij}(s) = I(Z_i(s^-) = j)$, $j = 1, 2$. Moreover, we let $X_{i1}(t)$ and $X_{i2}(t)$ be column vectors of external time-dependent covariates (Kalbfleisch and Prentice, 2011) and $X_i(s) = (X'_{i1}(s), X'_{i2}(s))'$. The history of the process is denoted $H_i(t) = \{(N_{i1}(s), N_{i2}(s)), X_i(s), 0 < s < t\}$. The complete intensity function for $j \to 3 - j$ transitions for individual i is

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N_{ij}(t) = 1 | H_i(t))}{\Delta t} = Y_{ij}(t) \lambda_{ij}(t | H_i(t)) \tag{2.2.1}$$

where $\Delta N_{ij}(t) = N_{ij}(t + \Delta t^-) - N_{ij}(t^-)$, $j = 1, 2$, $i = 1, 2, \ldots, m$.

If interest lies in events occurring over the period $(0, A]$, where A is a common administrative censoring time, let $C_i^\dagger$ denote a random censoring time assumed to be independent of the multistate process. We then let, $C_i = \min(A, C_i^\dagger)$, denote the net censoring time, $Y_i(s) = I(s \leq C_i)$, and $\bar{Y}_{ij}(s) = Y_i(s)Y_{ij}(s)$. We let $\bar{N}_{ij}(t) = \int_0^t \bar{Y}_{ij}(s)dN_{ij}(s)$ for $j = 1, 2$, $\bar{N}_i(t) = (\bar{N}_{i1}(t), \bar{N}_{i2}(t))$, and $\{\bar{N}_i(s), 0 < s\}$ denote the observed bivariate counting process. We assume that censoring can occur at any state. We, here, define $W_{i,N_{i2}(C_i)+1} = C_i - S_{iN_{i1}(C_i)}$ as a censoring time if censoring occurs during exacerbation; otherwise $W_{i,N_{i2}(C_i)+1} = 0$. The complete history of the observation and event processes is then $\bar{H}_i(t) = \{\bar{N}_i(s), Y_i(s), X_i(s), 0 \leq s < t\}$. We assume that two transitions cannot occur at the same time.

Let $U_i = (U_{i1}, U_{i2})'$ denote a bivariate individual specific random effect and consider a conditional intensity given $U_i = u_i$ as

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{ij}(t) = 1 | \bar{H}_i(t), u_i)}{\Delta t} = \bar{\lambda}_{ij}(t) | \bar{H}_i(t), u_i).$$

Under independent censoring (Kalbfleisch and Prentice, 2011) we can write

$$\bar{\lambda}_{ij}(t) | \bar{H}_i(t), u_i) = \bar{Y}_{ij}(t) \lambda_{ij}(t | H_i(t), u_i).$$

The role of the random effect here is to accommodate heterogeneity in the risk and the duration of events and to account for the dependence between the two processes.

Multiplicative semiparametric models can be expressed as follows. For the conditionally Markov intensity for $1 \to 2$ transition corresponding to the onset of symptoms, we let

$$\lambda_{i1}(t | H_i(t), u_{i1}) = u_{i1} Y_{i1}(t) d\Lambda_{01}(t) \exp(x_{i1}'(t)\beta_1).$$

For the conditionally semi-Markov intensity for $2 \to 1$ transitions representing resolution of symptoms, we let

$$\lambda_{i2}(t | H_i(t), u_{i2}) = u_{i2} Y_{i2}(t) d\Lambda_{02}(B_i(t)) \exp(x_{i2}'(t)\beta_2)$$

with $B_i(t) = t - S_{N_{i1}(t^-)}$ represents the time since symptom onset. Note that the terms $\Lambda_{0j}(\cdot)$ are infinite dimensional functions, but we informally write $\theta_j = (\Lambda_{0j}(\cdot), \beta_j')'$, for $j = 1, 2$ and let $\theta = (\theta_1', \theta_2')'$.

A bivariate random effects distribution can be formed through the use of a copula function $C(v_1, v_2)$, a bivariate distribution function with uniform $[0, 1]$ margins (Joe, 1997). The copula enables us to link any pair of marginal random effect distributions to obtain a bivariate distribution. The Gaussian copula is widely used and has the form $C(v_1, v_2; \rho) = \Phi_\rho(\Phi^{-1}(v_1), \Phi^{-1}(v_2))$ where $\Phi_\rho(\cdot)$ is a joint cumulative distribution function of a bivariate normal random variable with mean (0,0)', variances of one, and correlation coefficient ρ. The Archimedean family of copulas (Genest and Rivest, 1993) include the Clayton copula $C(v_1, v_2; \rho) = (v_1^{-\rho} + v_2^{-\rho} - 1)^{-1/\rho}$ which is widely used in survival analysis. If we let $G_j(u_j; \sigma_j)$ denote the marginal c.d.f for U_j for $j = 1, 2$, then

$$P(U_1 < u_1, U_2 < u_2; \phi) = G(u; \phi) = C(G_1(u_1; \phi_1), G_2(u_2; \phi_2); \rho) \tag{2.2.2}$$

denote the bivariate c.d.f indexed by $\phi = (\phi_1, \phi_2, \rho)'$ (Nelsen, 2006); we then let $\psi = (\theta', \phi')'$. The bivariate density function can then be written as

$$dG(u; \phi) = g_1(u_1; \phi_1) g_2(u_2; \phi_2) \frac{\partial^2 C(v_1, v_2; \rho)}{\partial v_1 \partial v_2}\bigg|_{(v_1, v_2) = (G_1(u_1; \phi_1), G_2(u_2; \phi_2))}.$$

where $g_j(u_j)$ is the marginal p.d.f for U_j for $j = 1, 2$. We consider both the Gaussian copula and the Clayton copula for which Kendall's τ is given by $2arcsin(\rho)/\pi$ and $\rho/(2+\rho)$, respectively. If there is a strong positive correlation between the two transitions, individuals at increased risk of $1 \to 2$ transition are at increased risk in $2 \to 1$ transitions (e.g. the resolution of exacerbation) which results in frequent short exacerbations. This pattern is often seen empirically. The severity of the chronic condition can be measured by the number of occurrence of excerbations; but this must be examined in concert with the sojourn time distribution.

The marginal likelihood for individual i is

$$\int_0^\infty \int_0^\infty P(Z_i(s), 0 < s < C_i | u_i, x_i(s), 0 < s < C_i; \theta) dG(u_i; \phi). \tag{2.2.3}$$

If we let

$$L_{i1}^{12} \propto \prod_{k=1}^{N_{i1}(C_i)} u_{i1} \lambda_{01}(s_{ik}) \exp(x_{i1}'(s_{ik})\beta_1) \exp\left(- \int_0^\infty u_{i1} \bar{Y}_{i1}(v) \exp(x_{i1}'(v)\beta_1) d\Lambda_{01}(v) \right)$$

and

$$L_{i1}^{21} \propto \prod_{k=1}^{N_{i2}(C_i)} u_{i2} \lambda_{02}(w_{ik}) \exp(x_{i2}'(s_{ik} + w_{ik})\beta_2) \exp\left(- \int_0^\infty u_{i2} \bar{Y}_{i2}(v) \exp(x_{i2}'(v)\beta_2) d\Lambda_{02}(B_i(v)) \right),$$

then $P(Z_i(s), 0 < s < C_i | u_i, x_i(s), 0 < s < C_i; \theta) = L_{i1}^{12} L_{i1}^{21}, i = 1, \ldots, m.$

Since $dG(u_i; \phi)$ cannot be factored, direct maximization of (2.2.3) is challenging. We adopt an expectation-maximization algorithm (Dempster *and others*, 1977) to facilitate semiparametric analyses.

Given the random effects we decompose the complete data log-likelihood into two parts as

$$l_C(\psi) = l_1(\theta) + l_2(\phi) \tag{2.2.4}$$

where

$$l_1(\theta) = \sum_{i=1}^m \{\log L_{i1}^{12} + \log L_{i1}^{21}\}$$
$$l_2(\phi) = \sum_{i=1}^m \log dG(u_i; \phi).$$

To implement the EM algorithm, we treat the random effects as missing data and the data on the event process as observed. Let $\psi^{(k)} = (\theta'^{(k)}, \phi'^{(k)})'$ denote the estimate of ψ at the kth iteration. In the E-step, we take the expectation of the complete log-likelihood (2.2.4)

with respect to u_i given the history $H_i(C_i)$. Let

$$Q(\psi;\psi^{(k)}) = Q_1(\theta;\psi^{(k)}) + Q_2(\phi;\psi^{(k)}) \tag{2.2.5}$$

where $Q(\psi;\psi^{(k)}) = \mathrm{E}[l_C(\psi)|H_i(C_i);\psi^{(k)}], Q_1(\theta;\psi^{(k)}) = \mathrm{E}[l_1(\theta)|H_i(C_i);\psi^{(k)}]$, and $Q_2(\phi;\psi^{(k)}) = \mathrm{E}[l_2(\phi)|H_i(C_i);\psi^{(k)}]$. Then evaluating (2.2.5) requires calculation of

i) $E[U_{ij}|H_i(C_i);\psi^{(k)}]$

ii) $E[\log U_{ij}|H_i(C_i);\psi^{(k)}]$

iii) $E[\log dG(U_i)|H_i(C_i);\psi^{(k)}]$.

For example,

$$E[U_{ij}|H_i(C_i);\psi^{(k)}] = \frac{\int_0^\infty \int_0^\infty u_{ij} P(H_i(C_i)|u_i, x_i(s), 0<s<C_i;\theta^{(k)}) dG(u_i;\phi^{(k)})}{\int_0^\infty \int_0^\infty P(H_i(C_i)|u_i, x_i(s), 0<s<C_i;\theta^{(k)}) dG(u_i;\phi^{(k)})} \tag{2.2.6}$$

where we write $P(H_i(C_i)|u_i, x_i(s), 0<s<C_i;\theta)$ for $P(Z_i(s), 0<s<C_i|u_i, x_i(s), 0<s<C_i;\theta)$ for convenience. The integrals in the numerator and denominator of (2.2.6) do not have closed forms so we use numerical integration by Gaussian-Quadrature with 32 nodes for each dimension. Here we exploited `OpenMP` in `C++` to make use of open multi-processing.

We let $\nu_{ij}^{(k)} = \log E[U_{ij}|H_i(C_i);\psi^{(k)}]$ at the kth iteration. At the $(k+1)$th M-step, we solve equations $U_{\beta_1}(\beta_1;\psi^{(k)}) = 0$ and $U_{\beta_2}(\beta_2;\psi^{(k)}) = 0$ for $\beta_1^{(k+1)}$ and $\beta_2^{(k+1)}$ where

$$U_{\beta_1}(\beta_1;\psi^{(k)}) = \sum_{i=1}^m \int_0^\infty \bar{Y}_{i1}(s)\left(x_{i1}(s) - \frac{\sum_{i=1}^m R_1^{(1,k)}(s;\beta_1)}{\sum_{i=1}^m R_1^{(0,k)}(s;\beta_1)}\right) dN_{i1}(s) \tag{2.2.7}$$

with

$$R_1^{(r,k)}(s;\beta_1) = \sum_{i=1}^m \bar{Y}_{i1}(s) x_{i1}^{\otimes r}(s) \exp(x_{i1}'(s)\beta_1 + \nu_{i1}^{(k)}), \quad r = 0, 1,$$

and

$$U_{\beta_2}(\beta_2;\psi^{(k)}) = \sum_{i=1}^m \sum_{j=1}^{N_{i2}(C_i)} \left(x_{i2}(s_{ij}+w_{ij}) - \frac{R_2^{(1,k)}(w_{ij};\beta_2)}{R_2^{(0,k)}(w_{ij};\beta_2)}\right) \tag{2.2.8}$$

with

$$R_2^{(r,k)}(w_{ij};\beta_2) = \sum_{h=1}^{m} \sum_{\ell=1}^{N_{h2}(C_h)+1} I(w_{h\ell} \geq w_{ij}) x_{h2}^{\otimes r}(s_{h\ell} + w_{ij}) \exp(x'_{h2}(s_{h\ell} + w_{ij})\beta_2 + \nu_{h2}^{(k)}),$$

for $r = 0, 1$.

Given $\hat{\beta}_j^{(k+1)}$, $j = 1, 2$, the estimates of the cumulative baseline intensities have the forms of Breslow profile likelihood estimates

$$\hat{\Lambda}_{01}^{(k+1)}(t) = \sum_{i=1}^{m} \int_0^t \frac{d\bar{N}_{i1}(s)}{\sum_{h=1}^{m} \bar{Y}_{h1}(s) \exp(x'_{h1}(s)\hat{\beta}_1^{(k)} + \nu_{h1}^{(k)})}$$

and

$$\hat{\Lambda}_{02}^{(k+1)}(w) = \sum_{i=1}^{m} \sum_{j=1}^{N_{i2}(C_i)} \frac{I(w = w_{ij})}{\sum_{h=1}^{m} \sum_{\ell=1}^{N_{h2}(C_h)+1} I(w_{h\ell} \geq w_{ij}) \exp(x'_{h2}(s_{hl} + w_{ij})\hat{\beta}_2^{(k)} + \nu_{h2}^{(k)})}$$

for the onset and duration of exacerbations respectively. The maximization of (2.2.5) in the semi-parametric setting can be easily carried out since $Q_1(\theta; \psi^{(k)})$ can be maximized using the `coxreg` function in `R` with $\nu_{ij}^{(k)}$ treated as an offset term. The estimate $\hat{\phi}^{(k+1)}$ is obtained by maximizing the auxiliary function

$$\sum_{i=1}^{m} E[\log dG(U_i; \phi) | H_i(C_i); \psi^{(k)}] \tag{2.2.9}$$

using a general optimization software such as the `optim` function in `R`. The E-step and M-step are repeated iteratively until the following stopping rule is satisfied:

$$\max(|\psi^{(k+1)} - \psi^{(k)}|) \leq \epsilon$$

where $\epsilon = 10^{-4}$ is used here.

To estimate standard errors, we use Louis's formula since the EM algorithm does not provide the observed information matrix directly (Louis, 1982). The observed information

matrix $I(\hat{\psi})$ is given as

$$I(\hat{\psi}) = \sum_{i=1}^{m} \left\{ -\mathrm{E}\left[\frac{\partial^2 l_C(\psi)}{\partial\psi\partial\psi'}\middle| H_i(C_i)\right] - VAR\left[\frac{\partial l_C(\psi)}{\partial\psi}\middle| H_i(C_i)\right]\right\}\Bigg|_{\psi=\hat{\psi}} \tag{2.2.10}$$

where $\hat{\psi}$ is the estimate of ψ. The score functions and the partial derivatives of score functions are provided in Appendix 2.A. The asymptotic distribution of estimators arising from a semiparametric model with Gamma frailty was investigated by Parner *and others* (1998) but the asymptotic distribution for bivariate frailty model requires development. We will show subsequently that the empirical performance is very good. Since the dimension of $I(\hat{\psi})$ increases as the number of individuals in the dataset increases, a convenient alternative is to use a model with piecewise constant baseline intensities.

2.3 Simulation Studies

Simulation studies were conducted to evaluate the finite sample performance of the estimators and the validity of the variance estimates from the joint analysis of Section 2.2 using the Gaussian copula, and the Clayton copula. We also examined the impact of model misspecification through the use of an independence model. The independence model is obtained by setting $U_{i1} \perp U_{i2}$; the model becomes the two independent frailty models.

We let $A = 2$ denote an administrative censoring time for individual i and generate data over the interval (0,2], We adopt an independent random censoring C_i which follows an exponential distribution with rate $-\log(0.9)/2$ giving 10% censoring *i.e.* $P(C_i < 2) = 0.1$. We consider the case of a fixed covariate and let x_i be the indicator of a randomized treatment where $x_i = 1$ if they are assigned to receive experimental treatment and $x_i = 0$ otherwise.

Under a time-homogeneous model, the intensity for the onset of exacerbations is $\lambda_{i1}(t|H_i(t), u_i) = u_{i1}\lambda_{01}\exp(x_i\beta_1)$ and the resolution of exacerbations is governed by a hazard function $\lambda_{i2}(t|H_i(t), u_i) = u_{i2}\lambda_{02}\exp(x_i\beta_2)$. We set $\lambda_{01} = 2$, $\lambda_{02} = 10$, $\beta_1 = \log(0.75)$, and $\beta_2 = \log(1.25)$. We let U_{ij} have log-normal or gamma distributions with $\mathrm{E}(U_{ij}) = 1$ and $\mathrm{Var}(U_{ij}) = \phi_j = 0.4$ for $j = 1, 2$, and we link u_{i1} and u_{i2} with the Gaussian copula or

the Clayton copula. Settings considered include with Kendall's $\tau = (-0.25, 0, 0.25)$ for the Gaussian copula and Kendall's $\tau = (0.25, 0.5)$ for the Clayton copula. We generate 1000 replicates in each scenario with sample sizes of $m = 500$ individuals.

Table 2.1, 2.2, and 2.3 reports results from the joint copula model and the misspecified independent frailty model with $U_{i1} \perp U_{i2}$. The empirical biases from the joint model are small, the empirical and average estimated standard errors of the parameters are in good agreement, and the empirical coverage probabilities are close to the nominal 95% level for all scenarios. From the independent frailty model, we observe that small biases in β_1 and β_2 and the empirical coverage probabilities are close to the nominal level. However, the frailty variance of resolution of exacerbation, ϕ_2, has non-negligible bias and the lower empirical coverage probabilities than the nominal 95% level, particularly when Kendall's τ is negative with the Gaussian copula and a strong positive dependence is present with the Clayton copula.

We also conducted additional simulation studies to assess the sensitivity of the analysis to misspecification of a copula function. The data were generated based on i) the Gaussian copula and ii) the Clayton copula with the gamma marginal distributions $\mathrm{E}(U_{ij}) = 1$ and $\mathrm{Var}(U_{ij}) = \phi_j = 0.4$ and the remaining parameters were unchanged. We use the Clayton copula for i), and the Gaussian copula for ii), respectively. Table 2.4 shows the results under this model misspecification. We note that the estimates of β_1, β_2 show little biases whereas ϕ_2 and Kendall's τ have biases. Therefore, if interest lies in the treatment effect on the onset of exacerbation or the resolution of exacerbation based on this limited study, the model provides a modest degree of robustness to misspecification of the copula function. The variance parameters and the dependence parameters for the two random effects, however, appear sensitive to misspecification of the copula function. This is in line with finding of others (McCulloch and Neuhaus, 2011). Note that one could also examine the empirical bias in estimate of the cumulative baseline Markov transition intensity for the onset of exacerbations as well as the cumulative hazard for the sojourn times in the exacerbation state. We do not do this here.

Table 2.1: Finite sample performance of estimators from semiparametric analyses of Markov/semi-Markov model under the Gaussian copula with log-normal margins and a (misspecified) independence model

		Gaussian Copula - LN Margin				Independence - LN Margin			
τ	PARAM[†]	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
-0.25	β_1	-0.002	0.080	0.078	0.947	0.003	0.080	0.078	0.941
	β_2	0.001	0.082	0.085	0.956	-0.013	0.084	0.084	0.939
	ϕ_1	-0.001	0.071	0.070	0.947	-0.018	0.068	0.068	0.925
	ϕ_2	-0.013	0.101	0.098	0.921	-0.053	0.088	0.089	0.862
	τ	-0.005	0.083	0.080	0.949				
0.00	β_1	-0.003	0.076	0.078	0.948	-0.002	0.078	0.078	0.951
	β_2	-0.001	0.084	0.085	0.944	-0.002	0.087	0.085	0.943
	ϕ_1	-0.003	0.070	0.068	0.931	0.000	0.069	0.068	0.947
	ϕ_2	-0.012	0.091	0.094	0.935	-0.011	0.090	0.093	0.940
	τ	0.003	0.081	0.079	0.945				
0.25	β_1	0.000	0.076	0.077	0.960	-0.002	0.076	0.078	0.950
	β_2	0.001	0.082	0.084	0.954	0.015	0.088	0.085	0.939
	ϕ_1	-0.002	0.068	0.067	0.940	0.007	0.069	0.067	0.938
	ϕ_2	-0.005	0.088	0.090	0.948	0.027	0.096	0.095	0.970
	τ	0.011	0.083	0.081	0.945				

[†] True values are $\beta_1 = -0.2877, \beta_2 = 0.2231, \phi_1 = 0.4, \phi_2 = 0.4$.

Table 2.2: Finite sample performance of estimators from semiparametric analyses of Markov/semi-Markov model under the Gaussian copula with gamma margins and a (misspecified) independence model

		Gaussian Copula - Gamma Margin				Independence - Gamma Margin			
τ	PARAM[†]	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
-0.25	β_1	-0.001	0.080	0.080	0.954	0.004	0.080	0.080	0.954
	β_2	0.000	0.091	0.094	0.962	-0.015	0.089	0.092	0.960
	ϕ_1	-0.004	0.054	0.054	0.944	-0.021	0.053	0.053	0.915
	ϕ_2	-0.010	0.065	0.065	0.944	-0.034	0.063	0.064	0.896
	τ	-0.001	0.080	0.078	0.946				
0.00	β_1	-0.001	0.079	0.080	0.954	-0.001	0.079	0.080	0.954
	β_2	0.002	0.090	0.092	0.956	0.002	0.091	0.092	0.954
	ϕ_1	-0.004	0.051	0.052	0.948	-0.004	0.051	0.052	0.947
	ϕ_2	-0.012	0.064	0.064	0.933	-0.012	0.064	0.064	0.935
	τ	0.005	0.077	0.079	0.948				
0.25	β_1	0.000	0.079	0.079	0.950	-0.004	0.079	0.080	0.951
	β_2	0.000	0.087	0.090	0.961	0.015	0.088	0.091	0.955
	ϕ_1	-0.004	0.052	0.051	0.945	0.003	0.052	0.052	0.950
	ϕ_2	-0.007	0.066	0.064	0.949	-0.001	0.065	0.063	0.947
	τ	0.011	0.079	0.080	0.956				

[†] True values are $\beta_1 = -0.2877, \beta_2 = 0.2231, \phi_1 = 0.4, \phi_2 = 0.4$.

Table 2.3: Finite sample performance of estimators from semiparametric analyses of Markov/semi-Markov model under the Clayton copula with gamma margins and a (misspecified) independence model

		Clayton Copula - Gamma Margin				Independence - Gamma Margin			
τ	PARAM[†]	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
0.25	β_1	0.002	0.079	0.079	0.953	-0.001	0.080	0.079	0.952
	β_2	-0.001	0.086	0.087	0.956	0.010	0.087	0.088	0.953
	ϕ_1	-0.003	0.050	0.051	0.950	-0.003	0.050	0.051	0.948
	ϕ_2	-0.013	0.066	0.067	0.932	-0.054	0.059	0.060	0.813
	τ	0.001	0.091	0.094	0.949				
0.50	β_1	-0.001	0.079	0.078	0.958	-0.006	0.079	0.079	0.954
	β_2	0.003	0.083	0.083	0.943	0.022	0.084	0.085	0.936
	ϕ_1	-0.001	0.053	0.051	0.940	-0.003	0.053	0.051	0.939
	ϕ_2	-0.011	0.066	0.068	0.933	-0.076	0.056	0.058	0.695
	τ	0.007	0.107	0.096	0.944				

[†] True values are $\beta_1 = -0.2877, \beta_2 = 0.2231, \phi_1 = 0.4, \phi_2 = 0.4$.

Table 2.4: Empirical results of semiparametric modeling under a misspecified copula model where data are generated by i) the Gaussian copula and ii) the Clayton copula. For i) the Clayton copula is used and for ii) the Gaussian copula is used for analysis; Kendall's $\tau = 0.25$

Analysis	i) Clayton Copula - Gamma Margin				ii) Gaussian Copula - Gamma Margin			
PARAM†	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
β_1	0.000	0.079	0.079	0.949	0.001	0.079	0.079	0.952
β_2	0.001	0.087	0.090	0.961	-0.001	0.086	0.087	0.957
ϕ_1	0.003	0.053	0.052	0.949	-0.008	0.050	0.051	0.942
ϕ_2	0.050	0.074	0.073	0.925	-0.059	0.060	0.060	0.804
τ	0.049	0.096	0.095	0.919	-0.053	0.078	0.081	0.904

† True values are $\beta_1 = -0.2877, \beta_2 = 0.2231, \phi_1 = 0.4, \phi_2 = 0.4, \tau = 0.25$.

2.4 Recurrent Hospitalization Among Individuals with Affective Disorder

To investigate the course of depressive or bipolar disorder, the Danish Psychiatric Central Research Register collected patients experiencing hospitalization with affective disorder over the period 1994-1999 (Kessing *and others*, 2004). A total of 10523 patients were recruited and individual data were recorded from the start of the first hospitalization. Kessing *and others* (2004) and Cook and Lawless (2013) analyzed this data set by stratification on the number of prior admissions or by incorporating it as covariate, and concluded an increased risk of admission with increasing numbers of prior hospitalizations. However, this does not fully explain the nature of hospitalization process since the duration of the hospitalization was not considered. We applied our proposed model to data from a psychiatric hospital re-admission studies to address this here.

Figure 2.2 displays the histogram of the numbers of hospitalizations experienced by

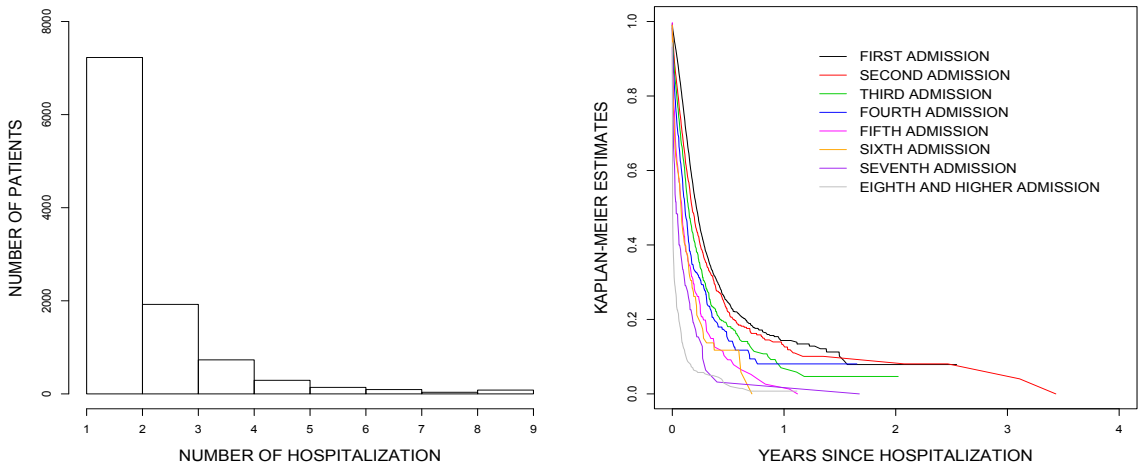


Figure 2.2: A crude summary of the distributions of the numbers of hospitalizations for individuals (left panel) and Kaplan–Meier estimates of the survivor functions for the duration of hospitalization stratified by the number of admissions (right panel) in the Danish Psychiatric Registry

individuals (panel a) and naive Kaplan–Meier estimates stratified by the number of prior admissions (panel b). The number of prior admissions were categorized into 1, 2, 3, 4, 5, 6, 7, and 8 or more prior admissions. The majority of patients experienced one or two admissions. The right panel in Figure 2.2 shows Kaplan-Meier estimates of the survivor functions $S_j(t) = P(W_{ij} \geq w)$ for $j = 1, \ldots, 7$ and $S_8(t) = P(W_{ik} \geq w)$ for $k \geq 8$, which indicates that the duration of hospitalizations tends to be shorter as the number of hospitalization increases. However, this crude summary of the data is misleading because it does not account for heterogeneity in the risk of admission or the duration of hospitalizations. As a consequence the Kaplan-Meier estimates are biased. While inverse probability of censoring weights can be used to correct for dependent gap times in the recurrent event setting (Lin *and others*, 1999; Cook and Lawless, 2007, Sec. 4.4.1), with alternating processes this is considerably more challenging and modeling the process offers a more convenient approach for dealing with this.

We consider the data from the admission of the first hospitalization. Gender, age at first diagnosis, type of disorder at the first discharge and cumulative number of admissions

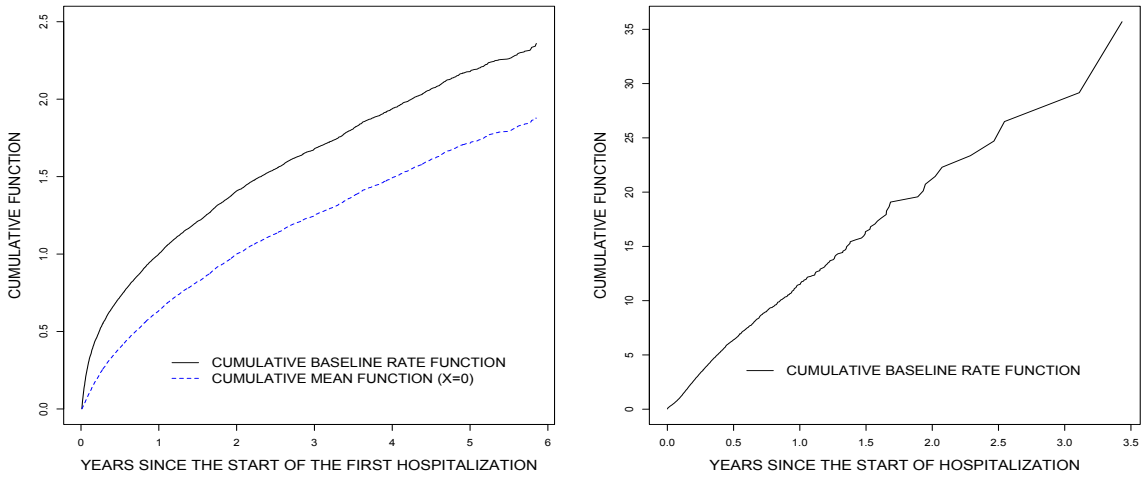
Table 2.5: Analysis of hospital re-admission data among individuals with affective disorder under a Gaussian copula with gamma margins

	Admission			Discharge		
Covariate	EST	SE	p	EST	SE	p
Sex : Female vs. Male	0.186	0.033	< 0.001	-0.184	0.021	< 0.001
Age at first hospitalization (Ref: $[0, 20)$)						
$[20, 40)$	-0.202	0.099	0.041	-0.065	0.066	0.325
$[40, 60)$	-0.382	0.098	< 0.001	-0.037	0.066	0.575
$[60, 80)$	-0.433	0.099	< 0.001	-0.095	0.066	0.150
$[80, \infty)$	-0.372	0.109	< 0.001	-0.141	0.072	0.050
Type of disorder (Ref: depression)						
Bipolar	0.299	0.043	< 0.001	-0.024	0.029	0.408
Cumulative number of admissions (Ref: 1)						
2	0.625	0.046	< 0.001	0.095	0.024	< 0.001
3	1.047	0.064	< 0.001	0.243	0.035	< 0.001
4	1.506	0.079	< 0.001	0.309	0.049	< 0.001
5	1.682	0.096	< 0.001	0.580	0.064	< 0.001
6	1.872	0.116	< 0.001	0.705	0.081	< 0.001
7	2.220	0.140	< 0.001	1.042	0.107	< 0.001
≥ 8	3.101	0.090	< 0.001	1.572	0.071	< 0.001
Frailty Variance	0.861	0.055		0.200	0.013	
Kendall's τ	-0.367	0.018	< 0.001			

were included as covariates. Age at first hospitalization was categorized as age $< 20, 20-40, 40-60, 60-80$ and 80 or over. The median of age at first hospitalization is 52 with range from 10 to 110. The objective of this analysis is to properly address the patients' cycle of hospitalization using an alternating two-state models and to identify the risk factors for the re-admission to a psychiatric hospital and the discharge.

The results of fitting the proposed model are reported in Table 2.5. We found that age at first discharge had a significant negative effect on the risk of re-admission, however, no significant effect on the rate of discharge from the hospital; older pateints have a lower rate

of re-admission and stayed longer in hospital. Patients with a diagnosis of bipolar disorder are at higher risk for re-admission (RR=1.35; 95% CI: $0.216, 0.38; p < 0.001$) and at lower risk for discharge from the hospital (RR=0.98; 95% CI: $-0.08, 0.03; p = 0.408$). This result is consistent with the findings of Cook and Lawless (2013). Sex had a significant impact on patients' cycle; men had a higher rate of re-admission (RR= 1.20; 95% CI: $0.12, 0.25; p < 0.001$) and a shorter duration of hospitalization (RR= 0.83, 95% CI: $-0.23, -0.14; p < 0.001$). Here we see a significant increasing trend in the risk of admission and discharge with increasing number of prior admissions. There are great subject-to-subject variations in the re-admission rate and the discharge rate among patients ($\hat{\phi}_1 = 0.861$, $\hat{\phi}_2 = 0.20$). The estimated Kendall's τ is -0.367, which means the transition times between the two states are negatively correlated. Therefore if an individual tends to have a higher rate of admission to hospital, the one has a lower rate of discharge from hospital leading to the longer duration of hospitalization.



(a) Admission to hospital (b) Discharge from hospital

Figure 2.3: Cumulative baseline rate function and cumulative mean function

The solid line in Figure 2.3 displays the estimated baseline cumulative hazard function of the time to re-admission and time to discharge for males with a depressive disorder aged [0-20). For the admission to hospital the two functions increases rapidly following the start

of the first hospitalization reflecting the risk of rapid re-admission. The mean function increases at a slower rate since individuals included in the risk set may still be admitted and hence not truly at risk of admission. Beyond this initial period the curves depart at a much slower rate in part due to the relatively short typical durations of admissions in relation to the total period of observation. Interest often lies in the marginal mean function for re-admissions given by

$$\mathrm{E}(N_{i1}(t)|x_i(t)) = \int_0^t \mathrm{E}(dN_{i1}(s)|x_i(s))$$

where

$$\mathrm{E}(dN_{i1}(s)|x_i(s)) = \int_0^\infty \int_0^\infty u_{i1} P(Y_{i1}(s) = 1|u_i, x_i(s)) d\Lambda_{01}(s) \exp(x'_{i1}(s)\beta_1) dG(u_i)$$

Since the estimation is based on the semi-parametric model, $P(Y_{i1}(s) = 1|u_i, x_i(s))$ is difficult to calculate; we obtain it via a simulation as follows. First, we generate a data set of 10,000 individuals based on the estimates obtained for each process upto the follow-up time. Due to the initial condition, the data were generated from the start of the first hospitalization. Next, we count the individuals who are in the symptom-free state and divide it by the total number of individuals at each time point. The dotted line in Figure 2.3 shows the estimated mean function for males with a depressive disorder aged [0-20). Clearly, it is less than the cumulative hazard function in which the mean function accounts for the risk set for the onset of re-admission.

Figure 2.4 contains plots of the baseline distributions for the durations of successive hospitalizations based on the fitted models. These are obtained by estimating the baseline hazard for the first 7 and then the 8th and subsequent hospitalizations and using these estimates to compute the corresponding semiparametric estimates of the survival functions. Compared to the right panel of Figure 2.3, the evidence of a trend is much lower from the fitted model, in part because the modeled and unexplained sources of heterogeneity have been accounted for.

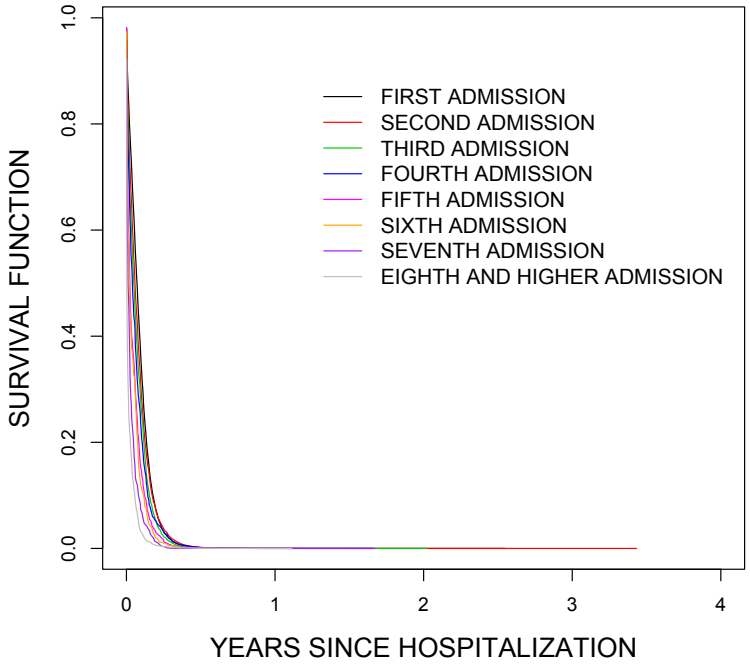


Figure 2.4: Survivor functions for the durations of successive hospitalizations based on the proposed model

2.5 Discussion

We have described a flexible bivariate random effect for the analysis of alternating two-state processes. This model is appealing when a long duration of episodes is observed and there is an association between the onset and the resolution of episodes. In particular, patients may receive treatments while experiencing episodes, which can affect subsequent relapses. In this setting, the accommodation of duration is sensible in recurrent data analysis. Also this model is more attractive since we can implement global hypothesis tests for treatment effects across two processes.

We viewed the recurrent episodes as an alternating two-state processes and formulated the intensity-based model in Section 2.2 using bivariate mixed models. The time since

onset of a chronic disease was used as the time-scale for the intensity governing the onset of exacerbations and a semi-Markov time scale was used for the intensity governing the resolution of exacerbations. The dependency between two processes is captured by introducing a copula function for random effects. The EM algorithm was conducted to obtain maximum likelihood estimators under the semi-parametric setting. The optimization of parameters for the variance of random effects and Kendall's τ is computationally intensive since the evaluation of (2.2.9) is required at every iteration during optimization.

Random effect models have a useful role when modeling multistate processes involving recurrent sojourns in one or more states when intensity-based analyses are not of interest (Putter and van Houwelingen, 2015). In the illustrative application insights are gained into the extent of heterogeneity and the nature of the dependencies. In settings where mortality rates are appreciable, such as in individuals with advanced chronic obstructive pulmonary disease, it may be of interest to generalize this model to incorporate a third state representing death; this would lead to a reversible illness-death model (Cook and Lawless, 2018). When estimating the cumulative mean function in this context, the terminal effect of death will naturally be accommodated. For the current model a difficulty arose in the estimation of $P(Y_{i1}(t) = 1|u_i, x_i(t))$ so we used resampling techniques to address this. This could be adopted in the context of a reversible illness-death model as well, but alternative approaches could also be investigated.

Appendix 2.A: The Score Functions and the Partial Derivatives of Score Functions

The observed information in (2.2.10) can be written as

$$I(\hat{\psi}) = \sum_{i=1}^{n} -\mathrm{E}\left[\frac{\partial^2 l_C(\psi)}{\partial\psi\partial\psi'}\middle| H_i(C_i)\right]\Big|_{\psi=\hat{\psi}} - \sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial l_C(\psi)}{\partial\psi}\frac{\partial l_C(\psi)}{\partial\psi}\middle| H_i(C_i)\right]\Big|_{\psi=\hat{\psi}} + \sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial l_C(\psi)}{\partial\psi}\middle| H_i(C_i)\right]\Big|_{\psi=\hat{\psi}}\mathrm{E}\left[\frac{\partial l_C(\psi)}{\partial\psi}\middle| H_i(C_i)\right]\Big|_{\psi=\hat{\psi}}$$

where $\frac{\partial l_C(\psi)}{\partial\psi} = \begin{pmatrix} \frac{\partial l_1(\theta)}{\partial\theta} \\ \frac{\partial l_2(\phi)}{\partial\phi} \end{pmatrix}$ and $-\frac{\partial^2 l_C(\psi)}{\partial\psi\partial\psi'} = \begin{pmatrix} -\frac{\partial^2 l_1(\theta)}{\partial\theta\partial\theta'} & 0 \\ 0 & -\frac{\partial^2 l_2(\phi)}{\partial\phi\partial\phi'} \end{pmatrix}$.

For simplicity, consider the case of a time-independent covariate vector.

The conditional score vector $\partial l_1(\theta)/\partial\theta$ for $\beta_1, \beta_2, d\Lambda_{01}(\cdot), d\Lambda_{02}(\cdot)$ has elements

$$\frac{\partial l_1(\theta)}{\partial\beta_1} = \sum_{i=1}^{m}\left\{N_{i1}(C_i)x_{i1} - u_{i1}\int_0^{\infty}\bar{Y}_{i1}(v)x_{i1}\exp(x_{i1}'\beta_1)d\Lambda_{01}(v)\right\},$$

$$\frac{\partial l_1(\theta)}{\partial d\Lambda_{01}(t_k)} = \frac{1}{d\Lambda_{01}(t_k)} - \sum_{i=1}^{m} u_{i1}\bar{Y}_{i1}(t_k)\exp(x_{i1}'\beta_1),$$

$$\frac{\partial l_1(\theta)}{\partial\beta_2} = \sum_{i=1}^{m}\left\{N_{i2}(C_i)x_{i2} - u_{i2}\int_0^{\infty}\bar{Y}_{i2}(v)x_{i2}\exp(x_{i2}'\beta_2)d\Lambda_{02}(B_i(v))\right\},$$

$$\frac{\partial l_1(\theta)}{\partial d\Lambda_{02}(w_k)} = \frac{1}{d\Lambda_{02}(w_k)} - \sum_{i=1}^{m}\sum_{j=1}^{N_{i2}(C_i)} u_{i2}I(w_{ij}\geq w_k)\exp(x_{i2}'\beta_2),$$

where t_k is the kth time of transition from $1\to 2$ and w_k is the kth gap time of transition

from $2 \to 1$.

The components of the conditional information matrix $-\partial^2 l_1(\theta)/\partial\theta\partial\theta'$ are as follows.

$$
\begin{aligned}
-\frac{\partial^2 l_1(\theta)}{\partial\beta_1\partial\beta_1'} &= \sum_{i=1}^{m} u_{i1} \int_0^{\infty} \bar{Y}_{i1}(v) x_{i1} x_{i1}' \exp(x_{i1}'\beta_1) d\Lambda_{01}(v), \\
-\frac{\partial^2 l_1(\theta)}{\partial\beta_1\partial d\Lambda_{01}(t_k)} &= \sum_{i=1}^{m} u_{i1} \bar{Y}_{i1}(t_k) x_{i1} \exp(x_{i1}'\beta_1), \\
-\frac{\partial^2 l_1(\theta)}{\partial\{d\Lambda_{01}(t_k)\}^2} &= \frac{1}{d\Lambda_{01}(t_k)^2}, \\
-\frac{\partial^2 l_1(\theta)}{\partial\beta_2\partial\beta_2'} &= \sum_{i=1}^{m} u_{i2} \int_0^{\infty} \bar{Y}_{i2}(v) x_{i2} x_{i2}' \exp(x_{i2}'\beta_2) d\Lambda_{02}(B_i(v)), \\
-\frac{\partial^2 l_1(\theta)}{\partial\beta_2\partial d\Lambda_{02}(w_{k)}} &= \sum_{i=1}^{m} \sum_{j=1}^{N_{i2}(C_i)} u_{i2} I(w_{ij} \geq w_k) x_{i2} \exp(x_{i2}'\beta_2), \\
-\frac{\partial^2 l_1(\theta)}{\partial\{d\Lambda_{02}(w_k)\}^2} &= \frac{1}{d\Lambda_{02}(w_k)^2}.
\end{aligned}
$$

The form of $l_2(\phi)$ depends on the marginal distribution of U_i, and the copula function. Note that for Gaussian copula with gamma margins or the Clayton copula with gamma margins, the score functions and the second derivatives of $l_2(\phi)$ with respect to ϕ do not have closed forms and numerical derivatives were obtained using the `grad` and `hessian` functions in `R`.

The conditional score vector and the components of the conditional information matrix for the Gaussian copula function with log-normal margins are present here. For convenience, we denote $\sigma_j = \text{Var}(\log(U_{ij}) = \sqrt{\log(\phi_j + 1)}$ and compute the conditional score vector and the conditional information matrix with respect to σ_j and obtain variance of ϕ_j using the delta method.

The log-likelihood function $l_2(\phi)$ is given as

$$
\begin{aligned}
l_2(\phi) = \sum_{i=1}^{m} \Bigg[& \frac{1}{2}\log(u_{i1}) - \frac{\log^2(u_{i1})}{2\sigma_1^2} - \log(\sigma_1) - \frac{\sigma_1^2}{8} + \frac{1}{2}\log(u_{i2}) - \frac{\log^2(u_{i2})}{2\sigma_2^2} - \log(\sigma_2) - \frac{\sigma_2^2}{8} \\
& - \frac{\log(1-\rho^2)}{2} - \frac{\rho^2(\sigma_1^2+\sigma_2^2)}{8(1-\rho^2)} + \frac{\rho\sigma_1\sigma_2}{4(1-\rho^2)} \\
& - \frac{\rho^2}{2(1-\rho^2)} \left\{ \frac{\log^2(u_{i1})}{\sigma_1^2} + \log(u_{i1}) + \frac{\log^2(u_{i2})}{\sigma_2^2} + \log(u_{i2}) \right\} \\
& + \frac{\rho}{1-\rho^2} \left\{ \frac{\log(u_{i1})\log(u_{i2})}{\sigma_1\sigma_2} + \frac{\sigma_2\log(u_{i1})}{2\sigma_1} + \frac{\sigma_1\log(u_{i2})}{2\sigma_2} \right\} \Bigg]
\end{aligned}
$$

The components of the conditional score vector $\partial l_2(\phi)/\partial\phi$ are given as follows.

$$
\begin{aligned}
\frac{\partial l_2(\phi)}{\partial \sigma_1} &= -\frac{m}{\sigma_1} - \frac{m\sigma_1}{4(1-\rho^2)} + \frac{m\rho\sigma_2}{4(1-\rho^2)} \\
&\quad + \sum_{i=1}^{m} \frac{\log^2(u_{i1})}{(1-\rho^2)\sigma_1^3} - \sum_{i=1}^{m} \frac{\rho\log(u_{i1})\log(u_{i2})}{(1-\rho^2)\sigma_1^2\sigma_2} + \frac{\rho}{1-\rho^2}\sum_{i=1}^{m} \left\{ \frac{-\sigma_2\log(u_{i1})}{2\sigma_1^2} + \frac{\log(u_{i2})}{2\sigma_2} \right\}, \\
\frac{\partial l_2(\phi)}{\partial \sigma_2} &= -\frac{m}{\sigma_2} - \frac{m\sigma_2}{4(1-\rho^2)} + \frac{m\rho\sigma_1}{4(1-\rho^2)} \\
&\quad + \sum_{i=1}^{m} \frac{\log^2(u_{i2})}{(1-\phi_0^2)\sigma_2^3} - \sum_{i=1}^{m} \frac{\phi_0\log(u_{i1})\log(u_{i2})}{(1-\phi_0^2)\sigma_1\sigma_2^2} + \frac{\rho}{1-\rho^2}\sum_{i=1}^{m} \left\{ \frac{-\sigma_1\log(u_{i2})}{2\sigma_2^2} + \frac{\log(u_{i1})}{2\sigma_1} \right\}, \\
\frac{\partial l_2(\phi)}{\partial \rho} &= \frac{m\rho}{1-\rho^2} - \frac{m\rho(\sigma_1^2+\sigma_2^2)}{4(1-\rho^2)^2} + \frac{m\sigma_1\sigma_2(1+\rho^2)}{4(1-\rho^2)^2} \\
&\quad - \frac{\rho}{(1-\rho^2)^2}\sum_{i=1}^{m} \left\{ \frac{\log^2(u_{i1})}{\sigma_1^2} + \frac{\log^2(u_{i2})}{\sigma_2^2} + \log(u_{i1}) + \log(u_{i2}) \right\} \\
&\quad + \frac{1+\rho^2}{(1-\rho^2)^2}\sum_{i=1}^{m} \left\{ \frac{\log(u_{i1})\log(u_{i2})}{\sigma_1\sigma_2} + \frac{\sigma_2\log(u_{i1})}{2\sigma_1} + \frac{\sigma_1\log(u_{i2})}{2\sigma_2} \right\}.
\end{aligned}
$$

The elements of the conditional information matrix $-\partial^2 l_2(\phi)/\partial\phi\partial\phi'$ are as follows.

$$
\begin{aligned}
-\frac{\partial^2 l_2(\phi)}{\partial\sigma_1\partial\sigma_1} &= -\frac{m}{\sigma_1^2} + \frac{m}{4(1-\rho^2)} + \sum_{i=1}^{m}\frac{3\log^2(u_{i1})}{\sigma_1^4(1-\phi_0^2)} - \sum_{i=1}^{m}\frac{2\rho\log(u_{i1})\log(u_{i2})}{(1-\rho^2)\sigma_1^3\sigma_2} - \frac{\rho}{1-\rho^2}\sum_{i=1}^{m}\frac{\sigma_2\log(u_{i1})}{\sigma_1^3},\\
-\frac{\partial^2 l_2(\phi)}{\partial\sigma_2\partial\sigma_2} &= -\frac{m}{\sigma_2^2} + \frac{m}{4(1-\rho^2)} + \sum_{i=1}^{m}\frac{3\log^2(u_{i2})}{\sigma_2^4(1-\rho^2)} - \sum_{i=1}^{m}\frac{2\rho\log(u_{i1})\log(u_{i2})}{(1-\rho^2)\sigma_1\sigma_2} - \frac{\rho}{1-\rho^2}\sum_{i=1}^{m}\frac{\sigma_1\log(u_{i2})}{\sigma_2^3},\\
-\frac{\partial^2 l_2(\phi)}{\partial\sigma_1\partial\sigma_2} &= -\frac{m\rho}{4(1-\rho^2)} - \sum_{i=1}^{m}\frac{\rho\log(u_{i1})\log(u_{i2})}{(1-\rho^2)\sigma_1^2\sigma_2^2} + \frac{\rho}{1-\rho^2}\sum_{i=1}^{m}\left\{\frac{\log(u_{i1})}{2\sigma_1^2} + \frac{\log(u_{i2})}{2\sigma_2^2}\right\},
\end{aligned}
$$

$$
\begin{aligned}
-\frac{\partial^2 l_2(\phi)}{\partial\sigma_1\partial\rho} &= \frac{m\rho\sigma_1}{2(1-\rho^2)^2} - \frac{m\sigma_2(1+\rho^2)}{4(1-\rho^2)^2} - \sum_{i=1}^{m}\frac{2\rho\log^2(u_{i1})}{(1-\rho)^2\sigma_1^3}\\
&\quad + \frac{(1+\rho^2)}{(1-\rho^2)^2}\sum_{i=1}^{m}\left\{\frac{\log(u_{i1})\log(u_{i2})}{\sigma_1^2\sigma_2} - \frac{\sigma_2\log(u_{i1})}{\sigma_1^2} + \frac{\log(u_{i2})}{2\sigma_2}\right\},\\
-\frac{\partial^2 l_2(\phi)}{\partial\sigma_2\partial\rho} &= \frac{m\rho\sigma_1}{2(1-\rho^2)^2} - \sum_{i=1}^{m}\frac{2\rho\log^2(u_{i2})}{(1-\rho)^2\sigma_2^3}\\
&\quad + \frac{(1+\rho^2)}{(1-\rho^2)^2}\sum_{i=1}^{m}\left\{\frac{\log(u_{i1})\log(u_{i2})}{\sigma_1\sigma_2^2} - \frac{\sigma_1\log(u_{i2})}{\sigma_2^2} + \frac{\log(u_{i1})}{2\sigma_2}\right\},\\
-\frac{\partial l_2(\phi)}{\partial\rho\partial\rho} &= -\frac{m(1+\rho^2)}{(1-\rho^2)^2} + \frac{(\sigma_1^2+\sigma_2^2)(1+3\rho^2)}{4(1-\rho^2)^3} - \frac{\sigma_1\sigma_2(3\rho+\rho^3)}{2(1-\rho^2)^3}\\
&\quad + \frac{1+3\rho^2}{(1-\rho^2)^3}\sum_{i=1}^{m}\left\{\frac{\log^2(u_{i1})}{\sigma_1^2} + \frac{\log^2(u_{i2})}{\sigma_2^2} + \log(u_{i1}) + \log(u_{i2})\right\}\\
&\quad - \frac{2\rho(3+\rho^2)}{(1-\rho^2)^3}\sum_{i=1}^{m}\left\{\frac{\log(u_{i1})\log(u_{i2})}{\sigma_1\sigma_2} + \frac{\sigma_2\log(u_{i1})}{2\sigma_1} + \frac{\sigma_1\log(u_{i2})}{2\sigma_2}\right\}.
\end{aligned}
$$

Chapter 3

Bias from Misspecified Semiparametric Rate-based Analysis of Recurrent Episodic Conditions

3.1 Introduction

In many chronic diseases individuals are at risk of recurrent episodic flares of symptoms. Statistical methods for recurrent event analyses have seen widespread application in such settings, including methods based on the semiparametric Andersen-Gill model (Andersen *and others*, 1993), marginal methods based on rate or mean functions (Lawless and Nadeau, 1995; Lin *and others*, 2000), and frailty models (Lawless, 1987; Klein, 1992; Wienke, 2010). These methods are geared towards the analysis of recurrent events which are instantaneous, but in many applications the events signal the onset of a symptomatic period (Hu *and others*, 2011) during which individuals are not at risk of an event. Examples include recurrent exacerbations of symptoms in individuals with chronic bronchitis (Grossman *and others*, 1998), recurrent bouts of depressive episodes in affective disorder (Kessing *and others*, 1999), and recurrent outbreaks of symptoms among individuals with herpes simplex virus infection (Romanowski *and others*, 2003). There are three approaches to how to handle these risk-free periods in the analyses of data from clinical trials. One may i) retain

individuals in the risk set for events, as is done in many medical studies, ii) simply remove individuals from the risk set while they are experiencing an episode, or iii) model the onset and duration times based on a two-state model. Alternating renewal processes were considered by Cox (1967) where the two types of sojourn times are assumed statistically independent. Several random effects (frailty) models have been developed to relax these independence conditions (Ng and Cook, 1997; Xue and Brookmeyer, 1996; Lee and Cook, 2018). Intensity-based two-state models offer a powerful framework for studying and describing the process dynamics, but since they require conditioning on the the process history they do not admit estimates of treatment effect with a causal interpretation. Although marginal methods can be robust to misspecification of the variance function or dependence structure, they cannot protect against misspecification of the risk set. Hu *and others* (2011) modeled the risk of recurrent hospitalizations, excluded individuals from the risk set during each admission, and fitted generalizations of the Prentice *and others* (1981) and Andersen and Gill (1982) models; since they were not interested in covariate effects on the duration of hospitalizations they did not model this feature. When interest lies in estimating the expected number of events over a period of follow-up or estimating the effects of associated covariates, it can be useful to model the probability of being in the symptom-free state which is obtainable from a more complete model for the onset and resolution of exacerbations. In clinical trials, however, intensity-based models do not offer a useful basis for causal inference when average treatment effects are the focus (Hernán and Robins, 2016).

Our objective is to study the asymptotic and empirical biases when standard recurrent event analyses (Andersen *and others*, 1993; Klein, 1992) are used in which individuals are considered in the risk set during episodes. Through specification of an alternating two-state process for the onset and resolution of exacerbations, we study the factors that lead to biases in semiparametric rate-based models. In the model which accommodates dependence in the risk for and duration of exacerbations we investigate the impact of the dependence on standard analyses with the adjusted risk set.

The remainder of this Chapter is organized as follows. In Section 3.2 we define notation and intensity functions for an alternating two-state process. In Section 3.3 we present the Andersen-Gill model, the associated estimating functions and the large sample results

for the estimated regression coefficients. The effect of misspecifying the risk set on the limiting behaviour of estimators is derived in Section 3.4 for both the one-sample problem and the regression setting. Section 3.4.1 considers the setting where the data are generated according to a homogeneous two-state model while Section 3.4.1 considers the case where there is heterogeneity in the risk for the onset and duration of exacerbations as well as dependence between associated random effects. The application of rate-based models is common in randomized clinical trials, so we study the implications on study power in Section 3.5. An application to a clinical trial involving individuals with herpes simplex virus infection is given in Section 3.6 and concluding remarks are made in Section 3.7.

3.2 Notation and an Alternating Two-state Model

Here we define a two-state which we consider as representing the underlying data generating process in order to study and characterize the factors which determine the biases in rate-based analyses.

Suppose individuals alternate between two states, a symptom-free state and a symptomatic state. Let $Z_i(s) = 1$ if individual i is symptom-free at $s > 0$, $Z_i(s) = 2$ if they are symptomatic, and suppose individuals start in state 1 at time t=0. Let $Y_{ij}(s) = I(Z_i(s^-) = j)$, $j = 1, 2$. We let S_{ik} and T_{ik} denote the onset time and resolution time of the kth exacerbation for individual i, and $W_{ik} = T_{ik} - S_{ik}$ be the duration of kth exacerbation $k = 1, \ldots$. The counting process $\{N_{ij}(u), 0 < u\}$ records the cumulative number of $j \to 3 - j$ transitions they experienced over $(0, t]$, $j = 1, 2$, where $N_{i1}(t) = \sum_{k=1}^{\infty} I(S_{ik} \leq t)$ records the cumulative number of onset times which are often the events of interest which interventions may be directed at preventing and $N_{i2}(t) = \sum_{k=1}^{\infty} I(T_{ik} \leq t)$ records the cumulative number of resolution times; $N_i(s) = (N_{i1}(s), N_{i2}(s))'$ is then a bivariate counting process. Let X_i be a set of fixed covariates, and $H_i(t) = \{N_i(s), 0 < s < t, X_i\}$.

Let C denote fixed administrative censoring time, $C_i^{\dagger}$ a random censoring time, and $C_i = \min(C, C_i^{\dagger})$. We assume that the censoring process is independent of the event process $\{N_i(s), 0 < s\}$, given covariates X_i. We let $Y_i(s) = I(s \leq C_i)$, $\bar{Y}_{ij}(s) = Y_i(s)Y_{ij}(s)$ and $\bar{N}_{ij}(t) = \int_0^t \bar{Y}_{ij}(s)dN_{ij}(s)$. If $\bar{N}_i(t) = (\bar{N}_{i1}(t), \bar{N}_{i2}(t))'$ then the complete history of the

observation and event processes is $\bar{H}_i(t) = \{\bar{N}_i(s), Y_i(s), 0 < s < t, X_i\}$. The complete intensity function for $j \to 3-j$ transitions for individual i can then be written as

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{ij}(t) = 1 | \bar{H}_i(t))}{\Delta t} = \bar{Y}_{ij}(t)\lambda_{ij}(t|H_i(t)), \qquad j = 1, 2, \tag{3.2.1}$$

under the condition of independence censoring (Kalbfleisch and Prentice, 2011).

Figure 3.1: A two-state diagram for chronic diseases with recurrent symptomatic episodes

Conditional on the censoring time, the probability of a particular sample path (Cook and Lawless, 2007) for individual i is

$$\prod_{k=1}^{N_{i1}(C_i)} \lambda_{i1}(s_{ik}|H_i(s_{ik})) \exp\left(-\int_0^{C_i} \bar{Y}_{i1}(u)\lambda_{i1}(u|H_i(u))du\right) \prod_{l=1}^{N_{i2}(C_i)} \lambda_{i2}(t_{il}|H_i(t_{il}))$$
$$\exp\left(-\int_0^{C_i} \bar{Y}_{i2}(u)\lambda_{i2}(u|H_i(u))du\right).$$

While likelihood based inference could be carried based on this specification our interest lies primarily in the settings of clinical trials where intensity-based analyses are undesirable. As pointed out by Kalbfleisch and Prentice (2011) conditioning on internal features of a life history process is undesirable when evaluating the effects of interventions. We emphasize therefore that the two-state model is to be used to derive limiting properties of estimators arising from marginal recurrent event analyses which are widely used in the clinical trial areas.

To study the potential bias of estimators arising from standard recurrent event analyses when there are risk-free periods, we consider two risk-set definitions (RSD) depicted in Figure 3.2. In RSD-A individuals are included in risk set for transitions from state 1 to state 2 event during symptomatic periods (i.e. $\bar{Y}_i^A(t) = Y_i(t)$); this represents a misspecification

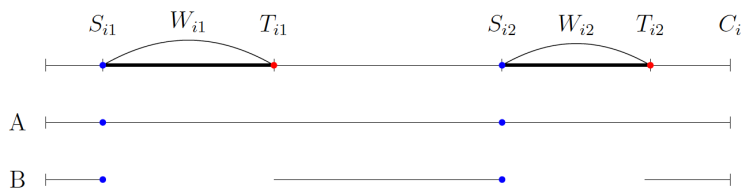


Figure 3.2: A schematic of a hypothetical timeline diagram with risk set definition (RSD) A and B

of the risk set in the sense that individuals are not truly at risk for the event (onset of episode) when they are in the midst of an episode an exacerbation. In RSD-B individuals are excluded from the risk during symptomatic periods (i.e. $\bar{Y}_i^B(t) = \bar{Y}_{i1}(t) = Y_i(t)Y_{i1}(t)$). In many ways RSD-B seems sensible since it is in alignment with how these periods would be treated in a multistate analysis. However in randomized clinical trials analyses are based on estimating marginal features, and exclusion of individuals from the risk set based on their status after randomization (which is possibly influenced by the treatment received) induces confounding and thereby compromising the ability to make causal statements (Cook and Lawless, 2018, Section 8.4). In the terminology of causal inference, whether an individual is at risk or not of an exacerbation is a collider in the causal path for effects of treatment on the onset of exacerbations (Cole *and others*, 2009; Hernán and Robins, 2016). In summary, excluding individuals from the risk set when they are experiencing an excerbation seems sensible since they truly are not at risk, but this precludes the ability to make direct causal statements about marginal features that may be of interest. Retaining individuals in the risk set is unnatural, but enables one to make a causal inference in a setting where the treatment effect has a natural interpretation. These points motivate us to explore the nature of biases induced when adopting the two approaches for defining the risk sets in marginal rate-based analyses, which we do in the context of a plausible underlying model for an alternating two-state process which we discuss in Section 3.4.

3.3 Standard Recurrent Event Analyses

3.3.1 The Semiparametric Andersen-Gill Model

The semiparametric Andersen-Gill model (Andersen and Gill, 1982) is based on a working Poisson model with a multiplicative covariate effect so that the rate function is given by

$$E(dN_{i1}(t)|H_i(t)) = E(dN_{i1}(t)|x_i) = dR_{01}(t)\exp(x_i'\gamma_1)$$

where the baseline rate function $dR_{01}(t)$ is not specified to have any particular parametric form. With a sample of m independent individuals the estimating functions for the AG model are

$$\sum_{i=1}^{m} \bar{Y}_i^A(t)\{dN_{i1}(t) - dR_{i1}(t)\} = 0 \tag{3.3.1}$$

$$\sum_{i=1}^{m} \int_0^\infty \bar{Y}_i^A(t)\{dN_{i1}(t) - dR_{i1}(t)\}x_{i1} = 0 \tag{3.3.2}$$

where $dR_{i1}(t) = dR_{01}(t)\exp(x_i'\gamma_1)$ and $R_{01}(t) = \int_0^t dR_{01}(s)ds$. Solving (3.3.1) with fixed γ_1 gives the profile "Breslow" estimate

$$d\tilde{R}_{01}^A(t;\gamma) = \frac{\sum_{i=1}^m \bar{Y}_i^A(t)dN_{i1}(t)}{\sum_{i=1}^m \bar{Y}_i^A(t)\exp(x_i'\gamma_1)}, \tag{3.3.3}$$

and substituting (3.3.3) into (3.3.2) gives an estimating function for γ_1 of the form

$$U^A(\gamma_1) = \sum_{i=1}^{m} \int_0^\infty \bar{Y}_i^A(s)\left\{x_i - \frac{\sum_{i=1}^m \bar{Y}_i^A(t)\exp(x_i'\gamma_1)x_i}{\sum_{i=1}^m \bar{Y}_i^A(t)\exp(x_{i1}'\gamma_1)}\right\}dN_{i1}(t). \tag{3.3.4}$$

We obtain $\hat{\gamma}_1^A$ by solving (3.3.4) and substitute $\hat{\gamma}_1^A$ into (3.3.3) to estimate $\hat{R}_{01}^A(t)$ as

$$\hat{R}_{01}^A(t) = \int_0^t \frac{\sum_{i=1}^m \bar{Y}_i^A(u)dN_{i1}(u)}{\sum_{i=1}^m \bar{Y}_i^A(u)\exp(x_i'\hat{\gamma}_1^A)}.$$

These results correspond to the setting for which the AG method was intended: where the events have no duration associated with them. Subject to the assumption of multiplicative covariate effects being correct this represents a valid analysis in such a setting. More generally (i.e. when the events signal the onset of a symptomatic period.), however, the properties of the resulting estimators have not been studied. If risk set definition B is used in an analysis, we proceed in the same fashion but replace $\bar{Y}_i^A(t)$ with $\bar{Y}_i^B(t)$ in (3.3.1) and (3.3.2).

3.3.2 Large Sample Robust Variance Formula and its Estimation

The estimating equations for the Andersen-Gill model are justified originally based on the working assumption that the events are generated by a Poisson process. Robust variance estimation is crucial, however, to provide protection from simple forms of model misspecification within the class of multiplicative rate-based models (Lawless and Nadeau, 1995; Lin *and others*, 2000). To obtain robust variance estimates, the model assumptions are relaxed to be simply $E(dN_{i1}(t)|x_i) = dR_{01}(t)\exp(x_i'\gamma_1)$. To accommodate different ways to define the risk set, we use $\bar{Y}_i^h(t)$ as the risk set indicator in the estimating equation (3.3.4) and write as the estimating function for γ_1^r

$$U^h(\gamma_1) = \sum_{i=1}^{m} \int_0^{\infty} \bar{Y}_i^h(t)\left\{x_{i1} - \frac{S^{(1,h)}(\gamma_1,t)}{S^{(0,h)}(\gamma_1,t)}\right\} dN_{i1}(t) \tag{3.3.5}$$

where $S^{(k,h)}(\gamma_1,t) = \sum_{i=1}^n \bar{Y}_i^h(t)\exp(x_i'\gamma_1)x_i^{\otimes k}$ for $k = 1, 2$, in which $a^{\otimes 2}$ means aa' and $a^{\otimes 1} = a$, and $a^{\otimes 0}$ represents a scalar 1. We let $\hat{\gamma}_1^h$ be the solution to $U^h(\gamma_1) = 0$, and let γ_1^h denote its limiting value which is determined by the solution to $E[U^h(\gamma_1)] = 0$ where the expectation is taken with respect to the true model. By adopting working model $E(dN_{i1}(t)|x_i, \bar{Y}_i^h(t) = 1) = dR_{01}(t)\exp(x_i'\gamma_1)$ following Lin *and others* (2000), (3.3.5) can be written as

$$U^h(\gamma_1) = \sum_{i=1}^{m} \int_0^{\infty} \left\{x_{i1} - \frac{S^{(1,h)}(\gamma_1,t)}{S^{(0,h)}(\gamma_1,t)}\right\} dM_{i1}^h(t)$$

where $dM_{i1}^h(t) = \bar{Y}_i^h(t)\{dN_{i1}(t) - dR_{01}(t)\exp(x_{i1}'\gamma_1)\}$. Since $\bar{Y}_i^h(t)$ is a predictable process (Andersen *and others*, 1993), $n^{-1/2}U^h(\gamma_1)$ is asymptotically $N(0, \mathcal{B}(\gamma_1))$ in distribution

where

$$\mathcal{B}(\gamma_1) = E\left[\left(\int_0^\infty \left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1, s)}{s^{(0,h)}(\gamma_1, s)}\right\} dM_{i1}^h(s)\right)\left(\int_0^\infty \left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1, t)}{s^{(0,h)}(\gamma_1, t)}\right\}' dM_{i1}^h(t)\right)\right],$$

and $s^{(k,h)}(t) = E[S^{(k,h)}(\gamma_1, t)]$ where here the expectation is taken with respect to the model given in Section 3.2. Since $n^{1/2}(\hat{\gamma}_1^h - \gamma_1^h) \simeq A^{-1}(\gamma_1^h) n^{-1/2} U(\gamma_1^h)$ by Taylor expansion, $n^{1/2}(\hat{\gamma}_1^h - \gamma_1^h)$ converges to $MVN(0, \mathcal{A}^{-1}(\gamma_1^h)\mathcal{B}(\gamma_1^h)\mathcal{A}^{-1}(\gamma_1^h))$ in distribution where

$$\mathcal{A}(\gamma_1) = E\left[\int_0^\infty \bar{Y}_i^h(t)\left\{\frac{s^{(2,h)}(\gamma_1, t)}{s^{(0,h)}(\gamma_1, t)} - \frac{s^{(1,h)}(\gamma_1, t)^{\otimes 2}}{s^{(0,h)}(\gamma_1, t)^2}\right\} dN_{i1}(t)\right].$$

The robust variance $\mathcal{A}^{-1}(\gamma_1^h)\mathcal{B}(\gamma_1^h)\mathcal{A}^{-1}(\gamma_1^h)$ is empirically estimated by $\hat{A}^{-1}(\hat{\gamma}_1^h)\hat{B}(\hat{\gamma}_1^h)\hat{A}^{-1}(\hat{\gamma}_1^h)$ in finite samples where

$$\hat{A}(\hat{\gamma}_1) = \frac{1}{m}\sum_{i=1}^m \left(\int_0^\infty \bar{Y}_i^h(t)\left\{\frac{S^{(2,h)}(\gamma_1, t)}{S^{(0,h)}(\gamma_1, t)} - \frac{S^{(1,h)}(\gamma_1, t)^{\otimes 2}}{S^{(0,h)}(\gamma_1, t)^2}\right\} dN_{i1}(t)\right)\Bigg|_{\gamma_1 = \hat{\gamma}_1^h},$$

$$\hat{B}(\hat{\gamma}_1) = \frac{1}{m}\sum_{i=1}^m \left(\int_0^\infty \left\{x_{i1} - \frac{S^{(1,h)}(\gamma_1, t)}{S^{(0,h)}(\gamma_1, t)}\right\} d\hat{M}_{i1}^h(t)\right)\left(\int_0^\infty \left\{x_{i1} - \frac{S^{(1,h)}(\gamma_1, t)}{S^{(0,h)}(\gamma_1, t)}\right\}' d\hat{M}_{i1}^h(t)\right)\Bigg|_{\gamma_1 = \hat{\gamma}_1^h},$$

and $d\hat{M}_{i1}^h(t) = \bar{Y}_i^h(t)\{dN_{i1}(t) - d\hat{R}_{01}^h(t)\exp(x_i\hat{\gamma}_1^h)\}$ where $d\hat{R}_{01}^h(t)$ is the estimate of $dR_{01}(t)$ based on the RSD-h for $h = A, B$. Then

$$\widehat{asvar}(n^{1/2}(\hat{\gamma}_1^h - \gamma_1^h)) = \hat{A}^{-1}(\hat{\gamma}_1^h)\hat{B}(\hat{\gamma}_1^h)\hat{A}^{-1}(\hat{\gamma}_1^h)$$

is used as a basis for inference.

3.4 Bias in Estimation of Mean Function and Regression Coefficients

Here we investigate the asymptotic bias of estimators under independent censoring when the marginal estimating equation is based on the Andersen-Gill model with a working in-

dependence assumption. We consider settings involving: a Markov/semi-Markov model for the onset and duration of recurrent exacerbations (Section 3.4.1), and a mixed model with dependent bivariate random effects modulating the baseline transition intensities (Section 3.4.2). Under each scenario, we obtain the asymptotic bias of the estimated mean function and the regression coefficients under the Andersen-Gill model.

3.4.1 Risk-set Misspecification in a Markov/Semi-Markov Model

When the intensity for the onset of exacerbations is based on a Markov model (3.2.1) reduces to the form

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{i1}(t) = 1 | \bar{H}_i(t))}{\Delta t} = \bar{Y}_{i1}(t)\lambda_{i1}(t), \tag{3.4.1}$$

where t is a total time (calendar time). If the resolution of exacerbations is governed by a semi-Markov model we obtain

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{i2}(t) = 1 | \bar{H}_i(t))}{\Delta t} = \bar{Y}_{i2}(t)\lambda_{i2}(B_i(t)) \tag{3.4.2}$$

where $B_i(t) = t - S_{N_{i1}(t^-)}$ is the time since symptom onset. For this specific investigation, we assume a time-homogeneous Poisson model for the onset of exacerbations and that exacerbation durations follow a Gamma distribution. Hu *and others* (2011) examined the asymptotic properties and convergence to the true value in the RSD-B setting so we focus on the asymptotic bias in RSD-A.

Marginal Rate and Mean Function Estimates

We consider the setting with no covariates first. In this case we have a single rate function estimate of interest $dR_{01}(t)$ and we consider a simplified version of (3.3.1) with a general at risk indicator $\bar{Y}_i^A(t)$:

$$\sum_{i=1}^{m} \bar{Y}_i^A(t)\{dN_i(t) - dR_{01}(t)\}$$

Taking the expectation of a single individual's contribution gives

$$E[\bar{Y}_i^A(t)(dN_{i1}(t) - dR_{01}(t))] = E[E[\bar{Y}_i^A(t)(dN_{i1}(t) - dR_{01}(t))|\bar{Y}_i^A(t)]]$$
$$= P(\bar{Y}_i^A(t) = 1)(E[dN_{i1}(t)|\bar{Y}_i^A(t) = 1] - dR_{01}(t)). \quad (3.4.3)$$

where the expectations and probabilities are computed based on the full model given in (3.3.1) under the assumptions in (3.4.1) and (3.4.2). We assume that the true transition intensity function $\lambda_{i1}(t)$ is time-homogeneous so that let $\lambda_{i1}(t) = \lambda_{01}$ and under completely independent censoring, equation (3.4.3) has solution

$$dR_{01}^A(t) = P(Y_{i1}(t) = 1)\lambda_{01}dt \quad (3.4.4)$$

and we note that $R_{01}^A(t) = \int_0^t dR_{01}^A(s)$ is the mean function for the counting process $\{N_{i1}(u), 0 < u\}$. The standard Nelson-Aalen estimator with RSD-A is consistent for the cumulative mean function under independent right censoring (Nelson, 1995; Lawless and Nadeau, 1995). It is therefore reasonable to use the Nelson-Aalen estimate when including individuals in the risk set to estimate the expected number of exacerbations. However, the estimator $\widehat{R}_{01}^A(t)$ will be asymptotically biased (conservative) for the true cumulative intensity (rate) function $\Lambda_{01}(t) = \int_0^t \lambda_{01}ds = \lambda_{01}t$. In other words, if $P(Y_{i1}(t) = 1)$ is small (i.e. if there is a high probability of being in the exacerbation state) the bias can get large. Since we assume that $W_{ik} \sim GAM(2, \lambda_{02})$ and using the fact that a Gamma random variable can be represented as a sum of independent exponential random variables we can use this to calculate $P(Y_{i1}(t) = 1)$. The details of the calculation of $P(Y_{i1}(t) = 1)$ is given in Appendix 3.A for a particular model with a Gamma distributed sojourn time distribution in state 2. We note that when $t \uparrow \infty, P(Y_{i1}(t) = 1)$ converges to $\lambda_{02}/(2\lambda_{01} + \lambda_{02})$ for the one sample problem. Figure 3.3 shows that the asymptotic bias of the Nelson-Aalen estimators decreases when the mean sojourn time in the exacerbation state decreases, in which case the effect of misspecification of the risk set is reduced.

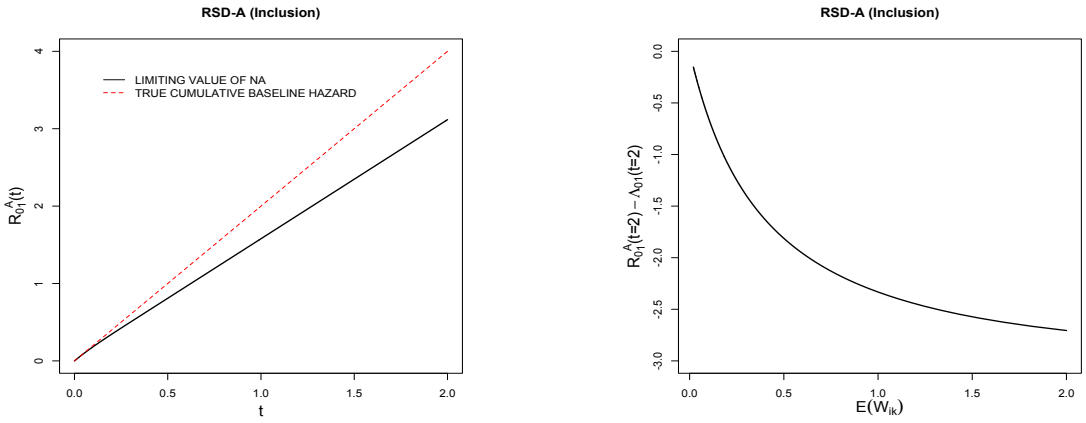


Figure 3.3: The limiting values and the asymptotic bias of Nelson-Aalen estimator under the RSD-A setting as a function of t with $E(W_{ik}) = 0.1$ (left panel) and as a function of $E(W_{ik})$ at t=2 (right panel) at fixed values of $\lambda_{01} = 2$, C=2, and 20% random censoring

Estimation in the Regression Setting

If we consider covariates, the expectation of the estimating equation (3.3.1) is given by

$$\int_{x_i} P(\bar{Y}_i^h(t) = 1|x_i)\{E^h[dN_{i1}(t)|\bar{Y}_i^h(t) = 1, x_i] - dR_{i1}(t)\}f(x_i)dx_i = 0 \tag{3.4.5}$$

where here this is taken with respect to the model given by (3.4.1) and (3.4.2). The proportional rate model with RSD-A does not account for the duration of exacerbations, so it is worthwhile to consider the asymptotic bias for $\hat{\gamma}_1^A$. In the sequel, we examine the limiting value, and asymptotic bias of $\hat{R}_{01}^A(t)$, and $\hat{\gamma}_1^A$.

Consider a randomized clinical trial where X_i is a binary variable with $P(X_i = 0) = P(X_i = 1) = 0.5$. We let $dR_{i1}^A(t) = dR_{01}^A(t)\exp(x_i\gamma_1^A)$ denote the value to which $d\hat{R}_{i1}(t)$ converges. Specifically we obtain

$$dR_{01}^A(t;\gamma_1^A) = \frac{\sum_{x=0}^1 P(x_i = x)P(Y_{i1}(t) = 1|x)\lambda_{01}\exp(x\beta_1)dt}{\sum_{x=0}^1 P(x_i = x)\exp(x\gamma_1^A)}. \tag{3.4.6}$$

The limiting value of $\hat{\gamma}_1^A$, denoted by γ_1^A can be obtained by solving

$$\int_0^\infty \left\{ s^{(1,A)}(u) - \frac{s^{(1,A)}(\gamma_1, u)}{s^{(0,A)}(\gamma_1, u)} s^{(0,A)}(u) \right\} du = 0, \tag{3.4.7}$$

where we define $s^{(k,A)}(u) = E[\bar{Y}_i^A(t) x_i^k dN_{i1}(u)]$, and $s^{(k,A)}(\gamma_1, u) = E[\bar{Y}_i^A(t) x_i^k \exp(x_i \gamma_1)]$ for $k = 1, 2$. Specifically,

$$s^{(0,A)}(t) = \sum_{x_i=0}^{1} P(x_i) P(\bar{Y}_{i1}(t) = 1 | x_i) \lambda_{01}(t) \exp(x_i \beta_1), \tag{3.4.8}$$

$$s^{(1,A)}(t) = P(x_i = 1) P(\bar{Y}_{i1}(t) = 1 | x_i = 1) \lambda_{01}(t) \exp(\beta_1), \tag{3.4.9}$$

$$s_1^{(0,A)}(\gamma_1, t) = \sum_{x_i=0}^{1} P(Y_i(t) = 1) P(x_i) \exp(x_i \gamma_1),$$

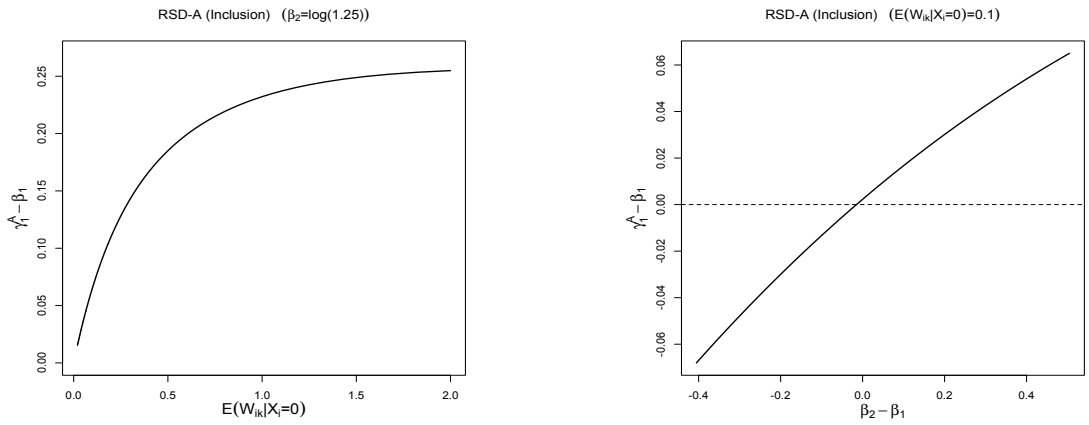
$$s_1^{(1,A)}(\gamma_1, t) = P(x_i = 1) P(Y_i(t) = 1) \exp(\gamma_1).$$

Note that $s^{(k,A)}(t) = s^{(k,B)}(t)$. If we solve (3.4.7), we obtain

$$\gamma_1^A = \beta_1 + \log \left(\frac{\int_0^\infty P(\bar{Y}_{i1}(u) = 1 | x_i = 1) du}{\int_0^\infty P(\bar{Y}_{i1}(u) = 1 | x_i = 0) du} \right), \tag{3.4.10}$$

where the detail of derivation is presented in Appendix 3.B. Here we assume the multiplicative model under the time-homogeneous assumption in (3.4.1) in which $\lambda_{i1}(t|x_i) = \lambda_{01} \exp(x_i' \beta_1)$, and we assume $W_{ik} \sim GAM(2, \lambda_{02} \exp(x_i' \beta_2))$ where W_{ik} is the duration of kth exacerbation. It suggests that in addition to the baseline functions, the magnitude of β_1 and β_2 determines the asymptotic bias of $\hat{\gamma}_1^A$ since $P(\bar{Y}_{i1}(t) = 1 | x_i)$ is a function of $\lambda_{i1}(t|x_i)$ and $\lambda_{i2}(t|x_i)$. The limiting value of robust covariance matrix is $\mathcal{A}^{-1}(\gamma_1^A) \mathcal{B}(\gamma_1^A) \mathcal{A}^{-1}(\gamma_1^A)$ the calculation of which is presented in Appendix 3.B.

Figure 3.4a shows that the asymptotic biases of coefficient decreases as the mean sojourn time for the exacerbation decreases under the setting $\lambda_{01} = 2, \beta_1 = \log(0.75)$, and $\beta_2 = \log(1.25)$ with C=2 and 20% random censoring. Even though the asymptotic bias becomes



(a) (b)

Figure 3.4: The asymptotic bias of a coefficient under the Andersen-Gill model with RSD-A as a function of $\mathrm{E}(W_{ik}|X_i = 0)$ (panel a), and $\beta_2 - \beta_1$ (panel b) at fixed values of $\lambda_{01} = 2$, and $\beta_1 = \log(0.75)$ with $C = 2$, and 20% random censoring

smaller, it does not converge to the true value. Figure 3.4b shows how the limiting value changes as a function of $\beta_2 - \beta_1$ in the setting where $\lambda_{01} = 2, \mathrm{E}(W_{ik}|X_i = 0) = 0.1$ ($\lambda_{02} = 20$), $\beta_1 = \log(0.75)$. The sign of bias for the coefficients depends on the difference between β_1 and β_2. If treatment significantly reduces the risk of the exacerbation and shortens the duration of symptoms, the misspecification of the risk set will lead to an underestimation of the treatment effect.

Simulation Studies

Simulation studies were conducted to examine the empirical bias and the performance of the robust variance estimator under misspecification of the risk set. To be specific the data are generated according to the two-state model in order for the data to represent that arising from episodic conditions. we are primarily interested in the performance of the AG type analyses with the original formulation using RSD-A and using the modified RSD-B to correspond to the common ad hoc approach for dealing with the duration of the episode and use of the AG model formulation. We set $\lambda_{01} = 2$, and

$\beta_1 = \log(0.75)$ where $\lambda_{i1}(t|x_i) = \lambda_{01}\exp(x_i\beta_1)$, $E[W_{ik}|x_i = 0] = 0.1$, 0.25, and 0.5 where $W_{ik}|x_i \sim GAM(2, \lambda_{02}\exp(x_i\beta_2))$. We also set C=2 and introduced 20% random censoring with a sample of size $m = 1000$ and a total of $nsim = 1000$ samples simulated. For individual i, we generate $S_{i1} \sim EXP(\lambda_{01}\exp(x_i\beta_1))$ followed by $T_{i1} = S_{i1} + W_{i1}$ where W_{i1} is generated by $GAM(2, \lambda_{02}\exp(x_i\beta_2))$. For $j > 1$, we generate $S_{ij}|S_{ij} > T_{ij}$ by truncated exponential distribution with rate $\lambda_{01}\exp(x_i\beta_1)$ and $W_{ik} \sim GAM(2, \lambda_{02}\exp(x_i\beta_2))$. We repeat this until either $S_{ij} > C_i$ or $T_{ij} > C_i$ for $j = 1, \ldots$ with a censoring time C_i. The results are reported in Table 3.1 under the RSD-A and RSD-B setting.

As expected from the asymptotic calculations the empirical bias of the estimated regression coefficients under RSD-A is positive when $\beta_1 < \beta_2$. This is because the positive effect of β_2 on the resolution of the exacerbation is reflected by the effect of β_1 resulting in $\hat{\beta}_1$ increasing. In other words, the average treatment effect under RSD-A in the spirit of causal inference is attenuated by the positive treatment effect for the resolution of disease. When $\beta_1 = \beta_2$, the empirical bias is low. Thus misspecification of the risk set is a significant matter if interest lies in the estimates of treatment effects. In a biological sense, the mechanism of occurrence and resolution of exacerbations may be different and it requires to reckon with a target of intrinsic treatment effects. Therefore, it needs caution to make a simple causal statement in this spirit. Interestingly, the use of a robust standard error induces lower empirical coverage probabilities than that of naive standard error since in this case the robust standard error is smaller than the naive standard error. From the formula of asymptotic variance in Appendix 3.C, when we use RSD-A, $\mathcal{A}(\gamma_1^A) = B_1^A$ and $B_3^A = B_4^A$, which means $\mathcal{B}(\gamma_1^B) = \mathcal{A}(\gamma_1^A) + B_2^A - B_4^A$. However, $B_2^A < B_4^A$ with RSD-A because $\mathrm{E}(dN_{i1}(s)dN_{i1}(t)|x_i, \bar{Y}_i^A(s) = 1, \bar{Y}_i^A(t) = 1)$ in B_2^A conditions that a subject should be at risk and in the exacerbation-free state at s and t where the exacerbation occurs at s, which implies that there is a transition from state 1 to 2 at s and another transition occurs from 2 to 1 between s^+ and t, whereas $R_{01}^A(s)$ and $R_{01}^A(t)$ in B_4^A only condition on being at risk and in the exacerbation-free state at s and t, separately. Thus, $\mathcal{A}(\gamma_1^A) > \mathcal{B}(\gamma_1^A)$ so that the naive standard error is greater than the robust standard error. Robust variance estimates ensure protection against model misspecification provided the rate function is correctly specified. However, if the risk set is misspecified inconsistent estimates are obtained and robust variance estimation does not provide protection against this formed

Table 3.1: Frequency properties of estimator from naive use of Andersen-Gill model; events simulated under two independent alternating processes with $\lambda_{i1}(t|H_i(t)) = \lambda_{01}\exp(x_i\beta_1)$ with $\lambda_{01} = 2, \exp(\beta_1) = 0.75, W_{ik} \sim GAM(2, \lambda_{02}\exp(\beta_2))$ over $(0, 2]$, sample size=1000, $nsim = 1000$

	RSD-A (Inclusion: $\bar{Y}_i^A(t) = Y_i(t)$)								RSD-B (Exclusion: $\bar{Y}_i^B(t) = Y_i(t)Y_{i1}(t)$)							
E$[W_{ik}\|x_i = 0]$	γ_1^A	AVE	SE^0	ESE	ASE^1	ASE^2	ECP^1	ECP^2	γ_1^B	AVE	SE^0	ESE	ASE^1	ASE^2	ECP^1	ECP^2
$\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 0.75$																
0.10	-0.285	-0.287	0.034	0.034	0.040	0.033	0.980	0.946	-0.288	-0.289	0.039	0.040	0.040	0.040	0.945	0.944
0.25	-0.277	-0.277	0.031	0.031	0.043	0.031	0.993	0.933	-0.287	-0.288	0.043	0.043	0.043	0.043	0.961	0.958
0.50	-0.258	-0.258	0.030	0.030	0.048	0.030	0.984	0.830	-0.288	-0.288	0.048	0.047	0.048	0.048	0.960	0.959
$\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 1.25$																
0.10	-0.222	-0.223	0.034	0.034	0.039	0.034	0.622	0.511	-0.288	-0.289	0.039	0.039	0.039	0.039	0.942	0.940
0.25	-0.160	-0.160	0.032	0.033	0.042	0.032	0.091	0.023	-0.288	-0.288	0.042	0.043	0.042	0.042	0.951	0.951
0.50	-0.103	-0.103	0.031	0.030	0.046	0.031	0.001	0.000	-0.288	-0.288	0.046	0.044	0.046	0.046	0.953	0.954

γ_1^h The limiting value of γ_1 for $h = A, B$.
[1] Naive results, ECP^1% is the empirical coverage probability for β_1 of a nominal 95% confidence intervals using the naive standard errors
[2] Robust results, ECP^2% is the empirical coverage probability for β_1 of a nominal 95% confidence intervals using the robust standard errors
[0] The limiting value of robust standard error

misspecification. As a result, the misspecification of risk set leads to significant biases and serious departure of empirical coverage probabilities. With RSD-B, the estimates of coefficients converge to the true value as indicated in Hu *and others* (2011).

3.4.2 Misspecification under Heterogeneity and Dependence

When there exists heterogeneity in the risk of exacerbation and the sojourn time distribution in the exacerbation state, a dependence between the two counting processes can be introduced. These features are typically ignored in recurrent event analyses so we further investigate the asymptotic bias of the Andersen-Gill estimators in this setting.

Suppose $U_i = (U_{i1}, U_{i2})'$ is a bivariate random effect for an alternating process so that under the assumption of independent censoring, the conditional intensity functions (3.2.1) take the form

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{i1}(t) = 1 | \bar{H}_i(t), u_i)}{\Delta t} = u_{i1} \bar{Y}_{i1}(t) \lambda_{i1}(t)$$

and

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{i2}(t) = 1 | \bar{H}_i(t), u_i)}{\Delta t} = u_{i2} \bar{Y}_{i2}(t) \lambda_{i2}(B_i(t))$$

where U_{ij} is gamma distributed with $\mathrm{E}(U_{ij}) = 1$ and $\mathrm{Var}(U_{ij}) = \phi_j$, for $j = 1, 2$ with the bivariate p.d.f $g(U_i)$.

Marginal Rate and Mean Function Estimates

In the absence of covariates, we assume $\lambda_{i1}(t) = u_{i1}\lambda_{01}$ and $W_{ik} \sim GAM(2, u_{i2}\lambda_{02})$. Then (3.4.4) becomes

$$dR^h_{01}(t) = E^h(U_{i1} | \bar{Y}^h_i(t) = 1) \lambda_{01} dt \tag{3.4.11}$$

for $h = A, B$, where $E^A[U_{i1} | \bar{Y}^A_i(t) = 1] = E_{U_i, Y_{i1}(t)=1 | Y_i(t)=1}[U_{i1} | Y_i(t) = 1]$ can be computed by

$$\int_0^\infty \int_0^\infty u_{i1} g(u_i, Y_{i1}(t) = 1 | Y_i(t) = 1) du_{i1} du_{i2} = \int_0^\infty \int_0^\infty u_{i1} g(u_i) P(Y_{i1}(t) = 1 | u_i) du_{i1} du_{i2},$$

and $E^B[U_{i1}|\bar{Y}_i^B(t)=1] = E_{U_i|\bar{Y}_{i1}(t)=1}[U_{i1}|\bar{Y}_{i1}(t)=1]$ is given by

$$\int_0^\infty \int_0^\infty u_{i1} g(u_i|\bar{Y}_{i1}(t)=1) du_{i1} du_{i2} = \int_0^\infty \int_0^\infty u_{i1} \frac{g(u_i)P(Y_{i1}(t)=1|u_i)}{P(Y_{i1}(t)=1)} du_{i1} du_{i2},$$

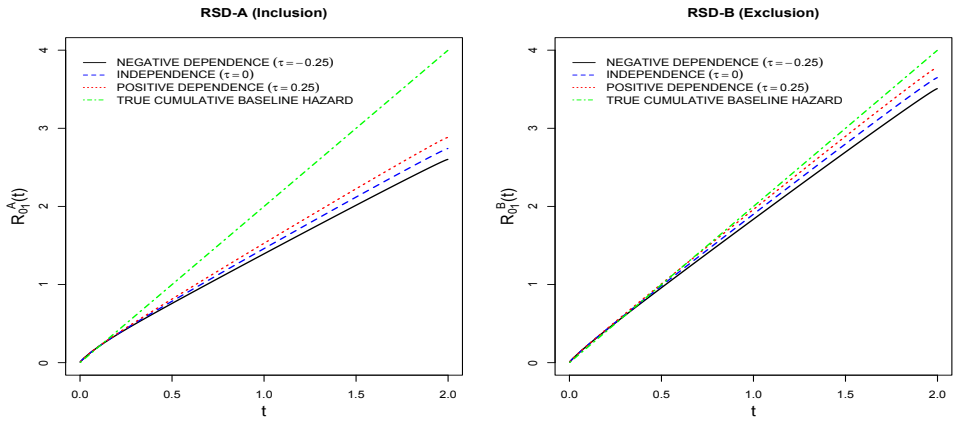
where

$$P(Y_{i1}(t)=1) = \int_0^\infty \int_0^\infty g(u_i)P(Y_{i1}(t)=1|u_i) du_{i1} du_{i2}.$$

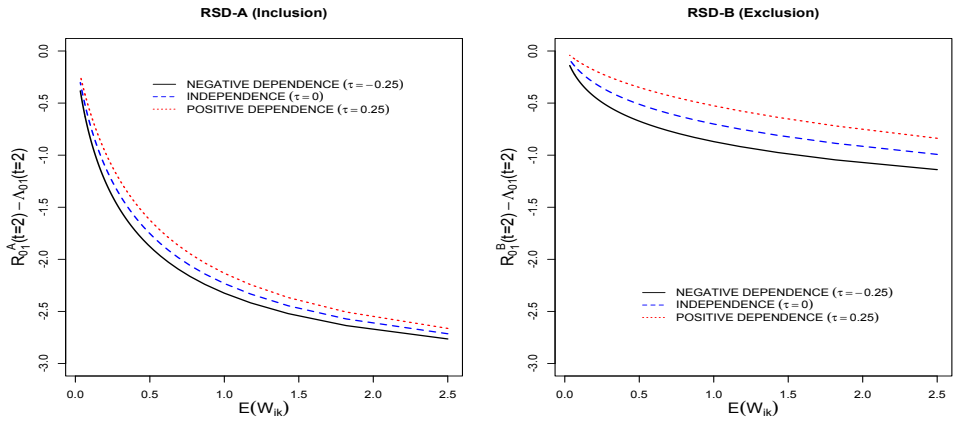
We obtain $P(Y_{i1}(t)=1|u_i)$ from the formula (3.A.2) or (3.A.3) in Appendix 3.A by replacing λ_{01} and λ_{02} with $u_{i1}\lambda_{01}$ and $u_{i2}\lambda_{02}$, respectively.

Figure 3.5a shows the limiting value of cumulative baseline function over the window (0,2] with 20% random censoring, $\lambda_{01} = 2, \mathrm{E}[W_{ik}|x_i = 0] = 0.25$. The true cumulative baseline hazard indicates $\Lambda_{01}(t) = \lambda_{01}t$. Figure 3.5b shows the asymptotic bias of cumulative baseline function at t=2 with 20% random censoring, $\lambda_{01} = 2$ as a function of the mean sojourn time in the exacerbation state. In both settings, we assume that U_{ij} is gamma distributed with mean 1 and variance $\phi_j = 0.4$ for $j = 1, 2$ and we link U_{i1} and U_{i2} with the Gaussian copula having Kendall's τ=-0.25, 0, and 0.25. Here the cumulative mean function is not equal to the cumulative intensity function due to symptom duration. We note that $R_{01}^A(t=2)$ and $R_{01}^B(t=2)$ are smaller than the true value of cumulative intensity function $\Lambda_{01}(t=2)$ and the bias also decreases as Kendall's τ increases; a strong positive association between U_{i1} and U_{i2} implies that the duration of exacerbations tends to decrease as the risk of exacerbations increases. As the mean sojourn time for exacerbations increases the bias increases in both RSD-A and RSD-B settings. Also, the use of RSD-B yields smaller bias than that of RSD-A. The Nelson-Aalen estimate with RSD-B shows a little departure from the true cumulative baseline hazard where the bias arises because of the impact of model misspecification in terms of individual heterogeneity and dependence between random effects for an alternating process.

Figure 3.5: The limiting value of the Nelson-Aalen estimate and the true cumulative baseline hazard under dependence sojourn time models due to correlated random effects



(a) Setting: $\lambda_{01} = 2, \mathrm{E}[W_{ik}|x_i = 0] = 0.25, \phi_1 = \phi_2 = 0.4$ with the Gaussian copula and $C = 2, r^{\dagger} = 20\%$



(b) Setting: $\lambda_{01} = 2, \phi_1 = \phi_2 = 0.4$ with the Gaussian copula and $C = 2, r^{\dagger} = 20\%$

Estimation in the Regression Setting

With covariates, we assume $\lambda_{i1}(t|u_{i1}, x_i) = u_{i1}\lambda_{01}\exp(\beta_1)$ and $W_{ik}|u_{i2}, x_i \sim GAM(2, u_{i2}\lambda_{02}\exp(\beta_2))$.

Then $dR_{i1}^h(t) = dR_{01}^h(t)\exp(x_i\gamma_1^h)$ can be expressed as

$$dR_{01}^h(t) = \frac{\sum_{x_i=0}^1 P(x_i)P(\bar{Y}_i^h(t)=1|x_i)E^h[U_{i1}|\bar{Y}_i^h(t)=1,x_i]\lambda_{01}\exp(x_i\beta_1)}{\sum_{x_i=0}^1 P(x_i)P(\bar{Y}_i^h(t)=1|x_i)\exp(x_i\gamma_1^h)}$$

for $h = A, B$ under a randomized clinical trial, where X_i is a binary variable with $P(X_i = 0) = P(X_i = 1) = 0.5$. The limiting value γ_1^h for $h = A, B$ can then be obtained by solving the equation (3.4.7) where

$$\begin{aligned}
s^{(0,h)}(t) &= \sum_{x_i=0}^1 P(x_i)P(\bar{Y}_i^h(t)=1|x_i)E^h[U_{i1}|\bar{Y}_i^h(t)=1,x_i]\lambda_{01}(t)\exp(x_i\beta_1),\\
s^{(1,h)}(t) &= P(x_i=1)P(\bar{Y}_i^h(t)=1|x_i=1)E^h[U_{i1}|\bar{Y}_i^h(t)=1,x_i=1]\lambda_{01}(t)\exp(\beta_1),\\
s_1^{(0,h)}(\gamma_1,t) &= \sum_{x_i=0}^1 P(x_i)P(\bar{Y}_i^h(t)=1|x_i)\exp(x_i\gamma_1),\\
s_1^{(1,h)}(\gamma_1,t) &= P(x_i=1)P(\bar{Y}_i^h(t)=1|x_i=1)\exp(\gamma_1),
\end{aligned}$$

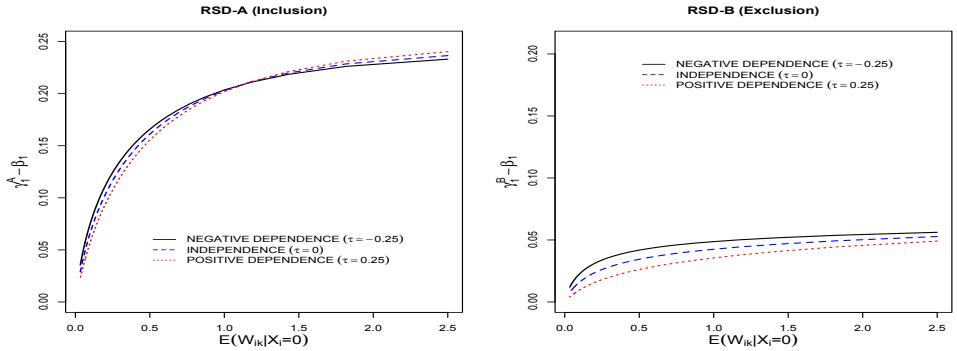
for $h = A, B$. Note that $s^{(k,A)}(t) = s^{(k,B)}(t)$. We can simplify γ_1^A here as

$$\gamma_1^A = \beta_1 + \log\left(\frac{\int_0^\infty P(\bar{Y}_i^A(u)=1)E^A[U_{i1}|\bar{Y}_i^A(u)=1,x_i=1]du}{\int_0^\infty P(\bar{Y}_i^A(u)=1)E^A[U_{i1}|\bar{Y}_i^A(u)=1,x_i=0]du}\right).$$

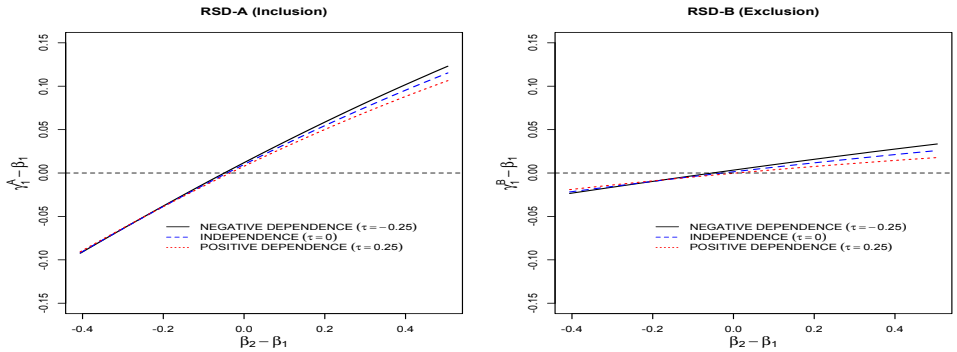
Likewise, the limiting value of the covariance estimator is given as $\mathcal{A}^{-1}(\gamma_1^h)\mathcal{B}(\gamma_1^h)\mathcal{A}^{-1}(\gamma_1^h)$ and the estimated variance is $\hat{A}^{-1}(\hat{\gamma}_1^h)\hat{B}(\hat{\gamma}_1^h)\hat{A}^{-1}(\hat{\gamma}_1^h)$ as shown in Section 3.3.2.

Figure 3.6 display the the asymptotic bias of regression coefficient γ^A, γ^B when the Andersen-Gill model is fitted. The bigger the mean sojourn time in the exacerbation state, the larger the bias is. For the second setting with $\beta_1 < 0$, the bias increases as β_2 is farther from β_1 and when $\beta_2 > \beta_1$, the bias is positive. Therefore, the effect of treatment for the risk of disease occurrence may be reduced if treatment decreases the occurrence of episodes and increases the recovery of episode under the Andersen-Gill model. Note that as Kendall's τ increases, the bias decreases.

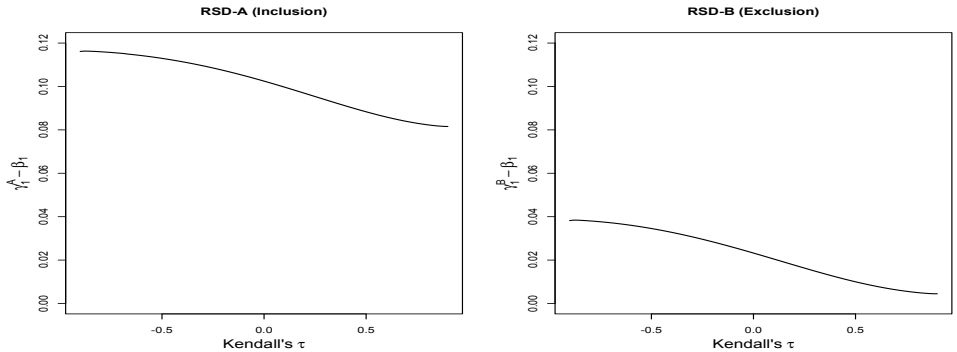
Figure 3.6: The asymptotic bias of γ^A and γ^B under the Andersen-Gill model for RSD-A and RSD-B with different β_2, $\mathrm{E}(W_{ik})$ and Kendall's τ



(a) Setting: $\lambda_{01} = 2, \beta_1 = \log(0.75)$, $\beta_2 = \log(1.25)$, $\phi_1 = \phi_2 = 0.4$, Kendall's $\tau = -0.25, 0, 0.25$ and $C = 2, r^\dagger = 20\%$



(b) Setting: $\lambda_{01} = 2, \mathrm{E}[W_{ik}|x_i = 0] = 0.25, \beta_1 = \log(0.75)$, $\phi_1 = \phi_2 = 0.4$, Kendall's $\tau = -0.25, 0, 0.25$ and $C = 2, r^\dagger = 20\%$



(c) Setting: $\lambda_{01} = 2, \mathrm{E}[W_{ik}|x_i = 0] = 0.25, \beta_1 = \log(0.75)$, $\beta_2 = \log(1.25), \phi_1 = \phi_2 = 0.4$, and $C = 2, r^\dagger = 20\%$

We conduct simulation studies to study the empirical bias and the performance of robust variance arising from the misspecification of risk set as well as frailty and dependence between alternating processes. The data is generated based on correlated random effects via a copula model. We set $\lambda_{01} = 2, \beta_1 = \log(0.75)$ where $\lambda_{i1}(t|H_i(t), u_i) = u_{i1}\lambda_{01}\exp(x_i\beta_1)$, $\mathrm{E}(W_{ik}|x_i = 0) = 0.1, 0.25$, and 0.5 where $W_{ik}|u_i, x_i \sim GAM(2, u_{i2}\lambda_{02}\exp(x_i\beta_2))$, and U_{ij} is gamma distributed with $\mathrm{E}(U_{ij}) = 1, \mathrm{Var}(U_{ij}) = 0.4$ for $j = 1, 2$. We use the Gaussian copula with Kendall's τ=-0.25, 0.00, and 0.25 for the bivariate distribution of U_i. We also set the administrative censoring time C=2, 20% random censoring with m=1000, and a total of 1000 samples were simulated. The results are reported in Table 3.2 under the RSD-A and RSD-B setting.

Table 3.2 shows that the means of estimated coefficients are almost equal to their limiting values. It is apparent that the impact of using an incorrect definition of the risk set can be appreciable, consistent with the result in Table 3.1. However, no concern of dependence between two alternating processes yields bias under the correct risk set RSD-B. As we observed in Figure 3.6, the bias decreases as Kendall's τ increases, the longer the mean sojourn time and the farther β_2 from β_1 the bigger the bias. There are differences between the naive standard errors and robust standard errors due to the model misspecification from the true model, but there is good agreement between the empirical standard error and the average robust standard error compared to the average naive standard error. Under the "correct" RSD-B, the robust variance estimates performed fairly well compared to the naive variance estimates although the empirical coverage probabilities are not in acceptable range when $E[W_{ik}|x_i = 0]$ is appreciable. However, the robust standard errors do not guarantee the protection against the misspecification of the risk indicator. When RSD-A is used, a serious bias and lower coverage probabilities are obtained in this setting. As a result it is important to take into account the duration of symptoms when specifying the risk set.

Table 3.2: Frequency properties of estimator from naive use of Andersen-Gill model; events simulated under conditional intensity-based model of Section 3.2 where $\lambda_{i1}(t|H_i(t), u_i) = u_{i1}\lambda_{01}\exp(x_i\beta_1)$ with $\lambda_{01} = 2, \exp(\beta_1) = 0.75, W_{ik}|x_i \sim GAM(2, \lambda_{02}\exp(x_i\beta_2))$, $\phi_1 = \phi_2 = 0.4, \tau = (-0.25, 0.00, 0.25)$ over (0,2] with 20% random censoring

	RSD-A (Inclusion: $\bar{Y}_i^A(t) = Y_i(t)$)								RSD-B (Exclusion: $\bar{Y}_i^B(t) = Y_i(t)Y_{i1}(t)$)							
τ	γ_1^A	AVE	SE^0	ESE	ASE^1	ASE^2	ECP^1	ECP^2	γ_1^B	AVE	SE^0	ESE	ASE^1	ASE^2	ECP^1	ECP^2
E$[W_{ik}\|x_i = 0] = 0.10$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 0.75$																
-0.25	-0.284	-0.284	0.048	0.046	0.041	0.048	0.921	0.956	-0.286	-0.287	0.055	0.052	0.041	0.055	0.881	0.958
0.00	-0.285	-0.283	0.050	0.050	0.040	0.050	0.886	0.953	-0.285	-0.287	0.056	0.055	0.040	0.056	0.845	0.959
0.25	-0.286	-0.283	0.052	0.051	0.039	0.052	0.870	0.955	-0.287	-0.284	0.056	0.055	0.039	0.056	0.829	0.960
E$[W_{ik}\|x_i = 0] = 0.25$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 0.75$																
-0.25	-0.276	-0.274	0.045	0.045	0.045	0.045	0.934	0.940	-0.284	-0.282	0.057	0.057	0.045	0.057	0.864	0.944
0.00	-0.278	-0.274	0.048	0.046	0.043	0.047	0.920	0.945	-0.286	-0.283	0.058	0.056	0.043	0.058	0.862	0.953
0.25	-0.280	-0.277	0.050	0.050	0.042	0.050	0.899	0.944	-0.288	-0.286	0.058	0.059	0.042	0.058	0.844	0.949
E$[W_{ik}\|x_i = 0] = 0.50$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 0.75$																
-0.25	-0.261	-0.259	0.043	0.043	0.049	0.044	0.939	0.912	-0.281	-0.278	0.060	0.059	0.049	0.060	0.885	0.955
0.00	-0.264	-0.262	0.046	0.046	0.048	0.046	0.928	0.913	-0.281	-0.285	0.061	0.062	0.048	0.061	0.869	0.944
0.25	-0.268	-0.265	0.048	0.047	0.046	0.048	0.910	0.925	-0.288	-0.286	0.061	0.060	0.046	0.061	0.873	0.955

γ_1^h The limiting value of γ_1 for $h = A, B$

[1] Naive results ECP^1% is the empirical coverage probability for β_1 of a nominal 95% confidence intervals using the naive standard errors

[2] Robust results, ECP^2% is the empirical coverage probability for β_1 of a nominal 95% confidence intervals using the robust standard errors

[0] The limiting value of robust standard error

Table 3.2 continued

	RSD-A (Inclusion: $\bar{Y}_i^A(t) = Y_i(t)$)								RSD-B (Exclusion: $\bar{Y}_i^B(t) = Y_i(t)Y_{i1}(t)$)							
τ	γ_1^A	AVE	SE^0	ESE	ASE^1	ASE^2	ECP^1	ECP^2	γ_1^B	AVE	SE^0	ESE	ASE^1	ASE^2	ECP^1	ECP^2
$E[W_{ik}\|x_i = 0] = 0.10$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 1.25$																
-0.25	-0.212	-0.211	0.049	0.047	0.040	0.049	0.516	0.646	-0.265	-0.263	0.055	0.052	0.040	0.054	0.810	0.940
0.00	-0.221	-0.219	0.051	0.050	0.040	0.051	0.555	0.727	-0.271	-0.269	0.056	0.055	0.040	0.056	0.818	0.941
0.25	-0.230	-0.223	0.053	0.051	0.038	0.052	0.617	0.807	-0.278	-0.275	0.056	0.055	0.038	0.056	0.824	0.948
$E[W_{ik}\|x_i = 0] = 0.25$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 1.25$																
-0.25	-0.163	-0.162	0.046	0.047	0.043	0.046	0.188	0.225	-0.254	-0.252	0.056	0.057	0.043	0.056	0.784	0.903
0.00	-0.171	-0.167	0.048	0.047	0.042	0.048	0.208	0.296	-0.261	-0.257	0.057	0.056	0.042	0.057	0.806	0.927
0.25	-0.180	-0.177	0.050	0.050	0.041	0.050	0.277	0.416	-0.270	-0.266	0.058	0.056	0.041	0.058	0.813	0.937
$E[W_{ik}\|x_i = 0] = 0.50$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 1.25$																
-0.25	-0.122	-0.120	0.044	0.045	0.047	0.044	0.050	0.037	-0.246	-0.244	0.059	0.060	0.047	0.059	0.762	0.880
0.00	-0.126	-0.126	0.046	0.046	0.046	0.046	0.059	0.064	-0.253	-0.252	0.060	0.059	0.046	0.059	0.804	0.912
0.25	-0.132	-0.131	0.048	0.049	0.045	0.048	0.086	0.113	-0.262	-0.261	0.060	0.060	0.045	0.060	0.814	0.924

γ_1^h The limiting value of γ_1 for $h = A, B$

[1] Naive results ECP^1% is the empirical coverage probability for β_1 of a nominal 95% confidence intervals using the naive standard errors

[2] Robust results, ECP^2% is the empirical coverage probability for β_1 of a nominal 95% confidence intervals using the robust standard errors

[0] The limiting value of robust standard error

3.5 Impact of the Episode Duration Distribution on Power

We previously examined the impact of misspecification of risk set on point estimates and the asymptotic variance of estimates. In this section we explore the impact of misspecification of the risk set on study power.

Consider the design of a randomized trial with recurrent responses. At the design stage, we assume a mixed Poisson model and calculate a sample size based on $\lambda_{i1}(t|u_{i1}, H_i(t)) = u_{i1}\lambda_{01}\exp(\beta_1 x_i)$ with $U_{i1} \sim Gamma(1/\phi_1, 1/\phi_1)$ (Cook and Lawless, 2007). If we want to test if an intervention has an effect on event occurrence we test $H_0 : \beta_1 = \beta_{10} = 0$ vs. $H_A : \beta_1 = \beta_{1A}$. However, the true underlying model requires one to consider the exacerbation duration. Hence interest lies in the change of power due to the duration of the exacerbation as well as the association between the occurrence of exacerbation and the recovery of exacerbation. As previous sections, we consider the Andersen-Gill model with a robust standard error with RSD-A and RSD-B for analysis and use the two-sided Wald test.

Suppose we consider the hypotheses $H_0 : \beta_1 = \beta_{10} = 0$ vs. $H_A : \beta_1 = \beta_{1A} = \log(0.75)$ under a two-sided test at the $\alpha_1 = 0.05$ level of significance and the power of $1 - \alpha_2 = 0.8$. Assume $\lambda_{01} = 2, \phi_1 = 0.4$ and a randomized trial gives $P(X_i = 1) = P(X_i = 0) = 0.5$. Then we can calculate the sample size under the assumption of mixed Poisson model given as

$$m \geq \left\{ \frac{z_{\alpha_1/2}\sqrt{asvar_0(\sqrt{m}(\hat{\beta}_1 - \beta_{10}))} + z_{\alpha_2}\sqrt{asvar_A(\sqrt{m}(\hat{\beta}_1 - \beta_{1A})}}{\beta_{1A}} \right\}^2, \tag{3.5.1}$$

where z_p represents the $(1-p)$-quantile for a standard normal distribution. $asvar_0(\cdot)$ and $asvar_A(\cdot)$ denote the asymptotic variance under the null and alternative hypotheses, respectively, where

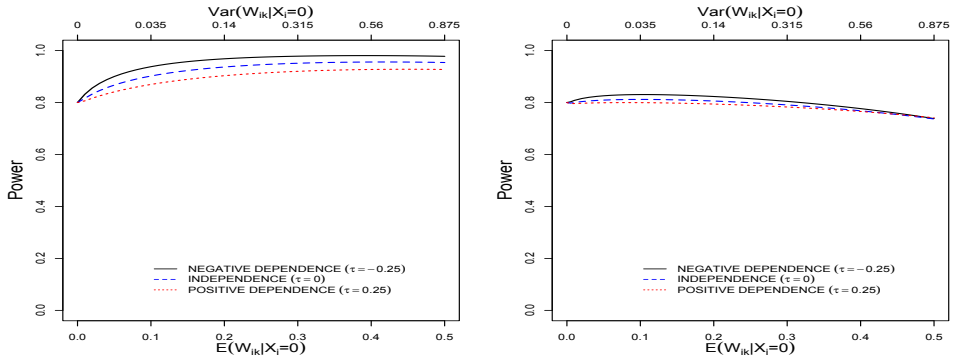
$$asvar(\sqrt{m}(\hat{\beta}_1 - \beta_1)) = \sum_{x_i=0}^{1} \left\{ P(X_i = x_i) E\left[\frac{\lambda_{01}\exp(\beta_1 x_i) C_i}{1 + \phi_1 \lambda_{01}\exp(\beta_1 x_i) C_i} \right] \right\}^{-1}.$$

Table 3.3: Empirical rejection rates for tests of treatment for occurrence of exacerbations where sample size is estimated based on the mixed Poisson model with $E[\bar{N}_{i1}(2)|x_i = 0] = 4$, $\phi_1 = 0.4$, $\beta_{10} = 0, \beta_{1A} = \log(0.75)$, E($W_{ik}|x_i = 0$)=0.10, 0.25, and 0.50, $\phi_2 = 0.4$ and Kendall's tau -0.25, 0, and 0.25 over (0,2] with 20% random censoring, $nsim = 2000$

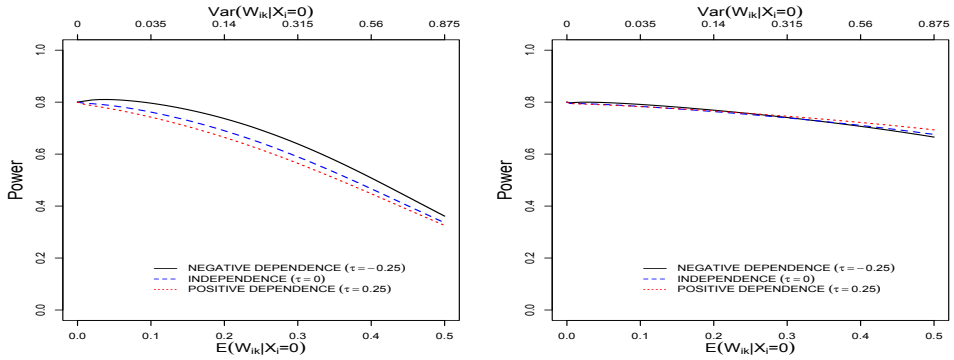
		τ=-0.25		$\tau = 0$		$\tau = 0.25$	
β_2	E($W_{ik}|x_i = 0$)	RSD-A	RSD-B	RSD-A	RSD-B	RSD-A	RSD-B
log(0.75)	0.10	92.1	81.5	88.2	79.4	85.8	79.1
	0.25	96.8	80.3	93.1	79.0	90.5	78.0
	0.50	97.2	72.4	95.4	72.3	92.3	73.2
0	0.10	77.4	77.8	73.5	77.1	72.6	77.0
	0.25	66.5	73.5	60.7	74.1	59.5	75.5
	0.50	34.5	64.4	34.2	67.7	30.9	67.8
log(1.25)	0.10	60.0	72.8	61.2	75.0	61.3	76.6
	0.25	28.3	68.7	29.7	69.5	30.6	73.1
	0.50	5.9	58.6	5.1	61.2	5.3	62.9

We next conduct simulation studies to investigate the impact of a misspecifed risk set and heterogeneity on power with the sample size calculated based on (3.5.1) where E$[\bar{N}_{i1}(2)] = 4$ with $\phi_1 = 0.4$ and 20% random censoring under the mixed Poisson model. We simulated data from a conditionally the Markov/semi-Markov model with correlated random effects via a copula model with the calculated sample size where E$[\bar{N}_{i1}(2)] = 4$ is fixed and $\phi_1 = \phi_2 = 0.4$. We consider different mean sojourn times for exacerbation state in the control arm of $E(W_{ik}|x_i = 0) = 0.1, 0.25, 0.5$ and different Kendall's τ=-0.25, 0, and 0.25.

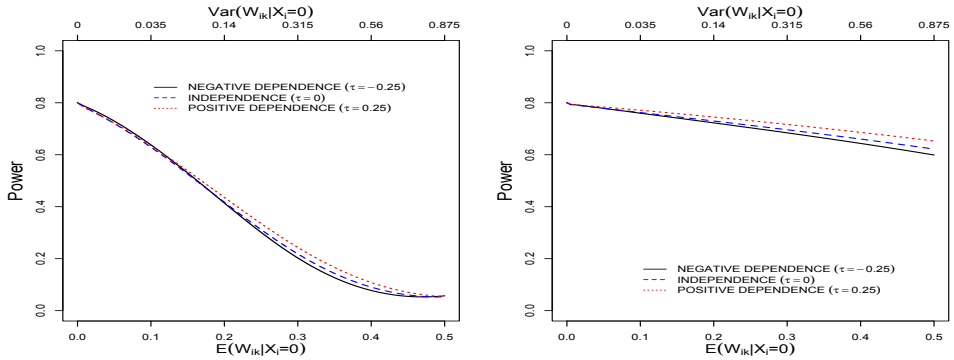
Figure 3.7: Power curves based on RSD-A (left panel) and RSD-B (right panel) with Kendall's τ -0.25, 0, and 0.25 where the sample size is calculated based on the mixed Poisson model with $E[\bar{N}_{i1}(2)|x_i = 0] = 4$, $\beta_{10} = 0, \beta_{1A} = \log(0.75)$, $\phi_1 = 0.4$, $\phi_2 = 0.4$



(a) $\beta_2 = \log(0.75)$



(b) $\beta_2 = 0$



(c) $\beta_2 = \log(1.25)$

With the generated data, we fit the Andersen-Gill model using RSD-A and RSD-B and conduct the hypothesis testing using robust standard errors based on a two-sided Wald test. A total number of 2000 replicates were generated and the empirical rejection rates (REJ%), defined as the percentage of replicates leading to rejection of the null hypothesis were computed; see Table 3.3.

The power decreases as the mean sojourn time for exacerbation-state increases for the settings with $\beta_1 \neq \beta_2$. It is worth noting that the power significantly depends on the value of β_2 with RSD-A; as β_2 is farther from the β_1 the loss in power increases. Note that when $\beta_1 = \beta_2$ we observe overpower with RSD-A. Using RSD-B reduces loss in power compared to RSD-A. As Kendall's τ increases there is a greater loss in power when $\beta_2 = 0$ under RSD-A. When $\beta_1 \neq \beta_2$, the increase in Kendall's τ decreases loss in power under the RSD-B.

Figure 3.7 shows power curves with the same setting as the empirical study. The effect of Kendall's τ on power relies on the mean sojourn time in the exacerbation-free state and the value of β_2. When $\beta_1 \neq \beta_2$, the increase in the mean sojourn time in the exacerbation-state reduces power, however, when $\beta_1 = \beta_2$ power is greater than 80% with RSD-A. The loss in power with RSD-B is smaller than the one with RSD-A when $\beta_1 \neq \beta_2$.

3.6 Application to a Herpes Simplex Trial

Herpes simplex is an infectious disease resulting in blisters on the infected part of the body. We consider the data from Romanowski *and others* (2003) who conducted a randomized two-period crossover trial to examine the effect of suppressive therapy versus episodic therapy. Here we only consider the first 24-week study period, therefore, each patients only had one treatment, suppressive or episodic therapy. We also include sex (female vs. male), virus type (HSV1 or HSV2) as covariates in addition to treatment (episodic therapy vs. suppressive therapy). In Table 3.4, we report on analyses of herpes simplex study using the Andersen-Gill model with RSD-A and RSD-B described in 3.3.1, respectively .

Table 3.4: Analysis of occurrence of herpes simplex using RSD-A and RSD-B based on the Andersen-Gill model

	RSD-A ($\bar{Y}_i^A(t) = Y_i(t)$)				RSD-B ($\bar{Y}_i^B(t) = Y_i(t)Y_{i1}(t)$)			
Covariate	EST	SE[1]	SE[2]	p[3]	EST	SE[1]	SE[2]	p[3]
Treatment	-1.875	0.200	0.240	< 0.001	-1.871	0.145	0.186	< 0.001
Sex	-0.189	0.135	0.173	0.276	-0.303	0.115	0.166	0.067
Virus Type	0.159	0.121	0.146	0.277	0.071	0.107	0.148	0.632

[1] Naive standard error

[2] Robust standard error

[3] p-values based on robust standard error

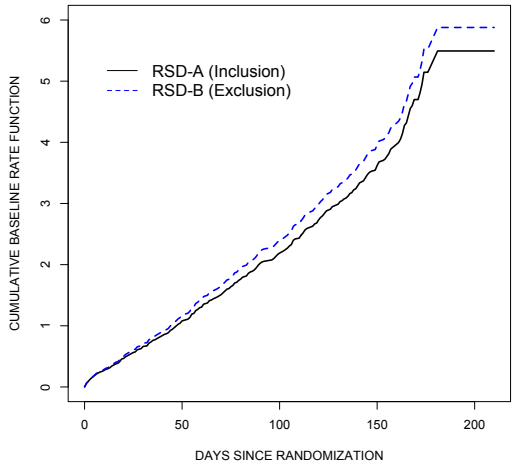


Figure 3.8: Cumulative baseline rate function with RSD-A (Inclusion) and RSD-B (Exclusion)

First of all, treatment has a significant effect on the occurrence of herpes simplex with both RSD-A ($RR = 0.15; 95\%$ CI: -2.35, -1.40; $p < 0.001$), RSD-B ($RR = 0.15; 95\%$

CI: -2.24, -1.51; $p < 0.001$), where two estimates are in very close agreement. Whereas the estimates for sex with RSD-A ($RR = 0.83; 95\%$ CI: -0.53, 0.15; $p = 0.276$) differ from the one with RSD-B ($RR = 0.74; 95\%$ CI: -0.63, 0.02; $p = 0.067$). Also there are differences in the estimates for virus type between RSD-A ($RR = 1.17; 95\%$ CI: -0.13, 0.45; $p = 0.277$) and RSD-B ($RR = 1.07; 95\%$ CI: -0.22, 0.36; $p = 0.632$). These differences arise due to the impact of misspecification of risk set. We also note that the naive standard errors and robust standard errors are not identical, which tells that there is heterogeneity between subjects and possible dependence between the onset and recovery of episode processes. Figure 3.8 contains a plot of the estimated cumulative baseline rate function with the Andersen-Gill analysis based on RSD-A and RSD-B. The slope of the cumulative baseline rate function with RSD-B is slightly greater than the one with RSD-A, suggestive of a higher rate for the occurrence of outbreaks with RSD-B than RSD-A (Cook and Lawless, 2007, Chapter 5, p. 177). However, the median of duration of episodes is 5 days indicating a short duration of episodes, therefore, the impact of misspecification of risk set on the estimate of cumulative baseline rate function is small. Note that with RSD-A, the cumulative baseline rate function can be naively interpreted as an estimate of the cumulative baseline mean function.

3.7 Discussion

In this Chapter we have pointed out that estimators of mean function and covariate effects from the naive use of the Andersen-Gill model (Andersen *and others*, 1993) are sensitive to the handling of risk-free periods as well as strength of the association between the onset and duration of episodic events. Misspecification of at risk indicators can lead to inconsistent estimators of regression coefficients and the use of robust standard errors does not guarantee protection against misspecification of the duration dependent processes. The biases we refer to for the mean function are specified in relation to the cumulative intensity for the onset of episodes, or the actual mean function reflected the expected number of events over time. In the regression setting we refer to the bias of estimators of the regression coefficient for the transition intensity for the onset of episodes.

Full specification of the intensities for an alternating two-state process is challenging in practice and it is impossible to achieve robustness in this framework since correct model specification is required to ensure the partial likelihood estimating equations are unbiased. Causal inference can be based on the expected number of events at a land-mark time or based on proportional rate function models but there is a tension between the need for full specification of models to advance scientific understanding and the need for simple models supporting causal conclusions. Lee and Cook (2018) develop a model for a mixed two-state process for characterizing recurrent episodic conditions which features a Markov time-scale for the onset of exacerbations and a semi-Markov time-scale for the duration of the exacerbations. Correlated random effects enable one to assess the need to accommodate heterogeneity and allow for a dependence between the sojourn times in the exacerbation state and the risk for the onset of events.

When mortality rates are appreciable, as is the case among individuals with advanced chronic obstructive pulmonary disease, it is considerably more challenging to model the onset and duration of exacerbations and summarize the effects of interventions. In the multistate framework an absorbing state representing death can be added, and random effects can be considered in the intensities for death. However expressing treatment effects robustly on the onset of exacerbations is very challenging. Much work has been carried out in this area for recurrent transient events (Cook and Lawless, 1997; Ghosh and Lin, 2000, 2002) but utility-based analyses may be preferable when events have a duration associated with them (Cook *and others*, 2003).

An alternative approach in these more complex settings is to focus on estimation of state occupancy probabilities using nonparametric methods. Cook and Lawless (2018, Sections 3.4 and 4.3) discuss this for one-sample problems and consider marginal regression models for state occupancy probabilities based on direct binomial regression (Scheike *and others*, 2008). Utility-based analyses are also of possible value (Cook *and others*, 2003, Cook and Lawless, 2018, Section 8.1). These and other marginal quantities, such as features of state entry time or sojourn time distributions, may offer a more convenient basis for causal inference since they are not defined inherently in terms of conditional probabilities. As always, the choice of the estimand must be made based on interpretation and it must be meaningful for the problem at hand. Inverse probability weighting can often be useful to

correct for some selection biases and confounding.

Appendix 3.A: Computation of State Occupancy Probabilities

$P(Y_{i1}(t) = 1|x_i)$ is difficult to calculate under the assumption of semi-Markov model, especially when the distribution of the duration of exacerbations is not exponential. Here, we decompose state 2 into two states to exploit the property of Gamma distribution which can be expressed as a sum of exponential distribution. We define a new state process $\{\bar{Z}(t), 0 < t\}$ on the extended state space $\{1, 2A, 2B\}$ (Cook *and others*, 2009), and let $Z(t) = 1$ if $\bar{Z}(t) = 1$ and $Z(t) = 2$ if $\bar{Z}(t) = 2A$ or $\bar{Z}(t) = 2B$ as shown in Figure 3.9. Then, $P(Y_{i1}(t) = 1|x_i)$ can be expressed by

$$\begin{aligned} P(Y_{i1}(t) = 1|x_i) &= P(Z(t) = 1|Z(0) = 1, x_i) \\ &= 1 - \sum_{r=2A,2B} P(\bar{Z}(t) = r|\bar{Z}(0) = 1, x_i) = P(\bar{Z}(t) = 1|\bar{Z}(0) = 1, x_i). \end{aligned}$$

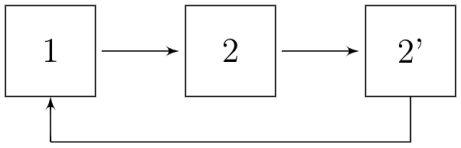


Figure 3.9: State diagram for recurrent exacerbations with extended Markov models

The term $P(\bar{Z}(t) = 1|\bar{Z}(0) = 1, x_i)$ is calculated by the transition probability matrix $\mathcal{P}(0, t|x_i) = \mathcal{P}(t|x_i) = [p_{ij}(t|x_i)]$, for $i, j = 1, 2A, 2B$. Here we consider the time-homogeneous case. We assume that the duration of the k^{th} exacerbation is $W_{ik} = W_{ik2A} + W_{ik2B}$ where $W_{ikl} \sim EXP(\lambda_{i2})$ for $l = 2A, 2B$ and $W_{ik2A} \perp W_{ik2B}$, so $W_{ik} \sim GAM(2, \lambda_{i2})$. Under the multiplicative model, we let $\lambda_{i1} = \lambda_{01} \exp(x_i\beta_1)$ and $\lambda_{i2} = \lambda_{02} \exp(x_i\beta_2)$. It is noted that there is a common covariate for the development and resolution of exacerbations. The time-homogeneous transition intensity matrix of $\{\bar{Z}(t), 0 < t\}$ on state space

$\{1, 2A, 2B\}$ is

$$Q = \begin{bmatrix} -\lambda_{i1} & \lambda_{i1} & 0 \\ 0 & -\lambda_{i2} & \lambda_{i2} \\ \lambda_{i2} & 0 & -\lambda_{i2} \end{bmatrix}.$$

We let $P_{12A}(t) = P(\bar{Z}(t) = 2A|\bar{Z}(0) = 1, x_i)$, $P_{12B}(t) = P(\bar{Z}(t) = 2B|\bar{Z}(0) = 1, x_i)$, and $P_{11}(t) = P(\bar{Z}(t) = 1|\bar{Z}(0) = 1, x_i)$. Using the Kolmogorov forward equations (Cox and Miller, 1965), we note

$$\begin{aligned} P'_{12A}(t) &= -\lambda_{i2}P_{12A}(t) + \lambda_{i1}P_{11}(t) \\ P'_{12B}(t) &= \lambda_{i2}P_{12A}(t) - \lambda_{i2}P_{12B}(t) \\ P'_{11}(t) &= \lambda_{i2}P_{12B}(t) - \lambda_{i1}P_{11}(t) \\ P_{12A}(t) + P_{12B}(t) + P_{11}(t) &= 1, \quad P_{11}(0) = 1 \end{aligned} \tag{3.A.1}$$

By solving the systems of equation of (3.A.1) we obtain $P_{11}(t)$ if the term $\lambda_{i1} - \lambda_{i2}/4 < 0$ as

$$P_{11}(t) = \frac{(\lambda_{i2})^2}{a^2 + b^2} + \exp(-at)\cos(bt)\left(\frac{2\lambda_{i1}\lambda_{i2}}{a^2 + b^2}\right) + \exp(-at)\sin(bt)\left(\frac{2\lambda_{i1}\lambda_{i2}(\lambda_{i2} - \lambda_{i1})}{(a^2 + b^2)2b}\right) \tag{3.A.2}$$

where $a = \lambda_{i1}/2 + \lambda_{i2}$ and $b = \sqrt{\lambda_{i1}\lambda_{i2} - (\lambda_{i1})^2/4}$. If $\lambda_{i1} - \lambda_{i2}/4 > 0$ it can be written as follows using Euler's formula,

$$\begin{aligned} P_{11}(t) = \frac{(\lambda_{i2})^2}{a^2 - (b')^2} &+ \exp(-at)\cosh(b't)\left(\frac{2\lambda_{i1}\lambda_{i2}}{a^2 - (b')^2}\right) \\ &+ \exp(-at)\sinh(b't)\left(\frac{2\lambda_{i1}\lambda_{i2}(\lambda_{i2} - \lambda_{i1})}{(a^2 - (b')^2)2b'}\right) \end{aligned} \tag{3.A.3}$$

where $b' = bi$. Likewise, if $\lambda_{i1} - \lambda_{i2}/4 < 0$, $P_{21}(t)$ is given as

$$P_{21}(t) = \frac{\lambda_{i2}}{2\lambda_{i1} + \lambda_{i2}} - \frac{\lambda_{i2}}{2\lambda_{i1} + \lambda_{i2}}\exp(-at)\cos(bt) - \frac{2\lambda_{i2}^2 + \lambda_{i1}\lambda_{i2}}{(2\lambda_{i1} + \lambda_{i2})2b}\exp(-at)\sin(bt)$$

else

$$P_{21}(t) = \frac{\lambda_{i2}}{2\lambda_{i1}+\lambda_{i2}} - \frac{\lambda_{i2}}{2\lambda_{i1}+\lambda_{i2}}\exp(-at)\cosh(b't) - \frac{2\lambda_{i2}^2+\lambda_{i1}\lambda_{i2}}{(2\lambda_{i1}+\lambda_{i2})2b'}\exp(-at)\sinh(b't)$$

Appendix 3.B: Calculation of Asymptotic Bias of $\hat{\gamma}_1^A$

We, here, derive $\gamma_1^A = \beta_1 + \log\left(\frac{\int_0^\infty P(\bar{Y}_{i1}(u)=1|x_i=1)du}{\int_0^\infty P(\bar{Y}_{i1}(u)=1|x_i=0)du}\right)$ in (3.4.10). By plugging $s^{(0,A)}(u)$, $s^{(1,A)}(u)$, $s^{(0,A)}(\gamma_1,u)$, and $s^{(1,A)}(\gamma_1,u)$ into (3.4.7), we have

$$\int_0^\infty \Bigg\{ P(\bar{Y}_{i1}(u)=1|x_i=1)\lambda_{01}\exp(\beta_1) - \frac{\exp(\gamma_1)}{1+\exp(\gamma_1)}\Bigg(P(\bar{Y}_{i1}(u)=1|x_i=1)\lambda_{01}\exp(\beta_1) + P(\bar{Y}_{i1}(u)=1|x_i=0)\lambda_{01}\Bigg)\Bigg\} du = 0$$

Then,

$$\frac{\exp(\gamma_1)}{1+\exp(\gamma_1)} = \frac{\int_0^\infty P(\bar{Y}_{i1}(u)=1|x_i=1)\lambda_{01}\exp(\beta_1)du}{\int_0^\infty \left(P(\bar{Y}_{i1}(u)=1|x_i=1)\lambda_{01}\exp(\beta_1) + P(\bar{Y}_{i1}(u)=1|x_i=0)\lambda_{01}\right)du} \tag{3.B.1}$$

We arrange (3.B.1) in terms of γ_1 so that

$$\exp(\gamma_1) = \exp(\beta_1)\frac{\int_0^\infty P(\bar{Y}_{i1}(u)=1|x_i=1)du}{\int_0^\infty P(\bar{Y}_{i1}(u)=1|x_i=0)du},$$

which has the final form as (3.4.10) by taking log() for both sides.

Appendix 3.C: Derivation of The Sandwich Covariance Matrix

Let $dM_{i1}^h(t) = \bar{Y}_i^h(t)\{dN_{i1}(t) - dR_{01}(t)\exp(x_{i1}\gamma_1)dt\}$. Then

$$\begin{aligned}
\mathcal{A}(\gamma_1) &= E\left[\int_0^\infty \bar{Y}_i^h(t)\left\{\frac{s^{(2,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)} - \frac{s^{(1,h)}(\gamma_1,t)^{\otimes 2}}{s^{(0,h)}(\gamma_1,t)^2}\right\}dN_{i1}(t)\right] \\
&= \sum_{x_i}\left[\int_0^C P(x_i)P(\bar{Y}_i^h(t)=1|x_i)\left\{\frac{s^{(2,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)} - \frac{s^{(1,h)}(\gamma_1,t)^{\otimes 2}}{s^{(0,h)}(\gamma_1,t)^2}\right\}\mathrm{E}(dN_{i1}(t)|x_i,\bar{Y}_i^h(t)=1)\right], \\
\mathcal{B}(\gamma_1) &= \mathrm{E}\left[\left(\int_0^\infty\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,s)}{s^{(0,h)}(\gamma_1,s)}\right\}dM_{i1}^h(s)\right)\left(\int_0^\infty\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)}\right\}dM_{i1}^h(t)\right)\right] \\
&= \mathrm{E}\left[\int_0^\infty\int_0^\infty\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,s)}{s^{(0,h)}(\gamma_1,s)}\right\}\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)}\right\}dM_{i1}^h(s)dM_{i1}^h(t)\right], \\
&= B_1^h + B_2^h - 2B_3^h + B_4^h \qquad (3.B.2)
\end{aligned}$$

where

$$\begin{aligned}
B_1^h &= \sum_{x_i}\int_0^C P(x_i)P(\bar{Y}_i^h(t)=1|x_i)\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)}\right\}^2\mathrm{E}(dN_{i1}^2(t)|x_i,\bar{Y}_i^h(t)=1), \\
B_2^h &= \int_0^C\int_0^C P(x_i)P(\bar{Y}_i^h(s)=1,\bar{Y}_i^h(t)=1|x_i)\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,s)}{s^{(0,h)}(\gamma_1,s)}\right\}\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)}\right\} \\
&\quad \times \mathrm{E}(dN_{i1}(s)dN_{i1}(t)|x_i,\bar{Y}_i^h(s)=1,\bar{Y}_i^h(t)=1), \\
B_3^h &= \int_0^C\int_0^C P(x_i)P(\bar{Y}_i^h(s)=1,\bar{Y}_i^h(t)=1|x_i)\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,s)}{s^{(0,h)}(\gamma_1,s)}\right\}\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)}\right\} \\
&\quad \times \mathrm{E}(dN_{i1}(s)|x_i,\bar{Y}_i^h(s)=1,\bar{Y}_i^h(t)=1)dR_{01}^h(t)e^{x_i\gamma_1},
\end{aligned}$$

and

$$\begin{aligned}
B_4^h &= \int_0^C\int_0^C P(x_i)P(\bar{Y}_i^h(s)=1,\bar{Y}_i^h(t)=1)\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,s)}{s^{(0,h)}(\gamma_1,s)}\right\}\left\{x_{i1} - \frac{s^{(1,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)}\right\} \\
&\quad \times e^{2x_i\gamma_1}dR_{01}(s)dR_{01}(t).
\end{aligned}$$

In Section 3.4.1 and under the assumption of time-homogeneous rate function for the two processes,

$$\mathrm{E}(dN_{i1}^2(t)|x_i, \bar{Y}_i^A(t) = 1) = \mathrm{E}(dN_{i1}(t)|x_i, \bar{Y}_i^A(t)) = P(Y_{i1}(t) = 1|x_i)\lambda_{01} \exp(x_i\beta_1),$$

and

$$\begin{aligned} &E(dN_{i1}(s)dN_{i1}(t)|x_i, \bar{Y}_i^A(s) = 1, \bar{Y}_i^A(t) = 1) \\ &\quad = P(\bar{Z}_i(s^-) = 1|\bar{Z}_i(0) = 1, x_i)P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(s) = 2, x_i)\lambda_{01}^2 \exp(2x_i\beta_1) \end{aligned}$$

for $s < t$, where $P(Y_{i1}(t) = 1|x_i) = P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(0) = 1, x_1)$, and $P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(s) = 2, x_i)$ is given in Appendix 3.A. In the setting of Section 3.4.2 with dependent random effects,

$$\begin{aligned} E(dN_{i1}^2(t)|x_i, \bar{Y}_i^A(t) = 1) &= \mathrm{E}(dN_{i1}(t)|x_i, \bar{Y}_i^A(t) = 1) \\ &= \int_0^\infty \int_0^\infty u_{i1} P(Y_{i1}(t) = 1|u_i, x_i)\lambda_{01}(t) \exp(x_i\beta_1) dG(u_i), \end{aligned}$$

and $E(dN_{i1}(s)dN_{i1}(t)|x_i, \bar{Y}_i^A(s) = 1, \bar{Y}_i^A(t) = 1)$ is given by

$$\int_0^\infty \int_0^\infty u_{i1}^2 P(\bar{Z}_i(s^-) = 1|\bar{Z}_i(0) = 1, x_i, u_i)P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(s) = 2, x_i, u_i)\lambda_{01}^2 \exp(2x_i\beta_1) dG(u_i)$$

for $s < t$. Moreover

$$E(dN_{i1}^2(t)|x_i, \bar{Y}_i^B(t) = 1) = \frac{\mathrm{E}(dN_{i1}^2(t)|x_i, \bar{Y}_i^A(t) = 1)}{P(Y_{i1}(t) = 1|x_i)},$$

and

$$E(dN_{i1}(s)dN_{i1}(t)|x_i, \bar{Y}_i^B(s) = 1, \bar{Y}_i^B(t) = 1) = \frac{\mathrm{E}(dN_{i1}(s)dN_{i1}(t)|x_i, \bar{Y}_i^A(s) = 1, \bar{Y}_i^A(t) = 1)}{P(Y_{i1}(s) = 1, Y_{i1}(t) = 1|x_i)}$$

for $s < t$, where $E(dN_{i1}(s)|x_i, \bar{Y}_i^B(s) = 1, \bar{Y}_i^B(t) = 1)$ is given by

$$\frac{\int_0^\infty \int_0^\infty u_{i1} g(u_i) P(\bar{Z}_i(s^-) = 1 | \bar{Z}_i(0) = 1, x_i, u_i) P(\bar{Z}_i(t^-) = 1 | \bar{Z}_i(s) = 2, x_i, u_i) du_{i1} du_{i2} \lambda_{01} \exp(x_i \beta_1)}{P(Y_{i1}(s) = 1, Y_{i1}(t) = 1 | x_i)}$$

for $s < t$ or

$$\frac{\int_0^\infty \int_0^\infty u_{i1} g(u_i) P(\bar{Z}_i(s^-) = 1 | \bar{Z}_i(0) = 1, x_i, u_i) P(\bar{Z}_i(t^-) = 1 | \bar{Z}_i(s) = 1, x_i, u_i) du_{i1} du_{i2} \lambda_{01} \exp(x_i \beta_1)}{P(Y_{i1}(s) = 1, Y_{i1}(t) = 1 | x_i)}$$

for $t < s$.

Chapter 4

Dependence Modeling for Multi-Type Recurrent Events Via Copulas

4.1 Introduction

4.1.1 Overview

In many chronic diseases individuals are at risk of several distinct types of potentially recurring events. In asthma, for example, individuals are at risk of different types of recurrent exacerbations (Jayaram *and others*, 2006), individuals with diabetes are at risk recurrent complications in eyes and kidneys (The Diabetes Control and Complications Trial Research Group, 1986), and cancer patients are at risk of metastases in different locations of the body (Hortobagyi, 1998). In public health studies there is interest in modeling the occurrence of different kinds of infections in populations at risk such as children in developing countries (Lemaire *and others*, 2011). A natural goal is to carry out a marginal analysis by estimating the rate of onset for each type of infection. However, the different types of infections may arise due to the same underlying risk factors (e.g. a compromised

immune system due to malnutrition, exposure to contaminated areas, etc.). It can therefore be informative to model the association between different types of infections.

Cai and Schaubel (2004) proposed semi-parametric marginal models for multi-type recurrent events data and develop robust standard errors. Chen *and others* (2012) developed additive marginal models, proved the consistency of estimators, and derived their asymptotic distribution under a working independence assumption. Cook *and others* (2010) proposed a copula-based bivariate mixed Poisson model with correlated random effects. Mazroui *and others* (2013) also considered bivariate frailty models for two types of recurrent events and death associated with two types of events based on maximum likelihood with a piecewise baseline hazard function and maximum penalized likelihood. Also (Cook and Lawless, 2007, Chapter 6) introduces different approaches in multitype recurrent events.

Even though the asymptotic theory is typically developed for multi-type recurrent event data for the general setting, applications provided typically deal with only two types of events. Frequentist methods based on flexible multivariate frailty models can be challenging to implement with more than 2 event types, particularly when semiparametric methods are of interest. We address this by developing a joint model for multiple types of recurrent events using a multivariate random effects distribution constructed using a copula model to link the component-specific random effects. This structure is appealing in that it enables separate modeling of heterogeneity and dependence and offers a natural basis for use of a composite "pairwise" likelihood approach to avoid the computational burden of the full likelihood. We also investigate an even more computationally convenient two-stage estimation procedure based on pairwise likelihood in which marginal models are fitted for each type of event at the first stage, and the dependence parameters are estimated at the second stage. Large sample theory is developed for both of these approaches.

The remainder of the Chapter is organized as follows. In the next sub-Section we provide a brief review of composite likelihood. In Section 4.2 we introduce notation, provide details on the model formulation, and give the full and composite likelihoods. An expectation-maximization algorithm is given in Section 4.3 for semiparametric analysis based on multiplicative rate function models; variance estimation is given in an Appendix 4.A. Section 4.4 reports on simulation studies investigating the finite sample properties of the simultaneous and two-stage estimation procedure based on pairwise composite likeli-

hood and an application is given in Section 4.5 on a motivating study on the effect of iron supplementation on the occurrence of different types of infections in malnourished children. Concluding remarks are made in Section 4.6.

4.1.2 Review of Composite Likelihood

We let θ denote a $p \times 1$ parameter of interest. In modeling multivariate data or in other settings involving complex dependence structures, the full likelihood may be complex or too computationally demanding to work with. As an alternative to full likelihood, Lindsay (1988) propose using a composite likelihood defined as

$$CL(\theta; y) = \prod_{j=1}^{J} L_j(\theta; y)^{w_j}, \tag{4.1.1}$$

a weighted product of marginal or conditional likelihood contributions $L_j(\theta; y)$; this may be viewed as an as the extension of the concept of pseudo-likelihood (Besag, 1974). Each term $L_j(\theta; y)$ is determined by the selection of $\{A_1, \ldots, A_J\}$, a set of marginal or conditional events, where $L_j(\theta; y) \propto f(y \in A_j; \theta)$ (Varin *and others*, 2011). Varin (2008) provided the excellent review of composite likelihood in different fields and classified composite likelihood contributions as based on conditional or marginal likelihoods. Composite conditional likelihoods are based on the product of conditional densities given conditions which the analyst specifies, whereas the latter is constructed from marginal densities.

As in ordinary likelihood, the composite likelihood score equations are unbiased estimating equations under mild regularity conditions. The maximum composite likelihood estimator $\hat{\theta}$, obtained by solving the $p \times 1$ equation

$$S(\theta) = \partial \log CL(\theta; y)/\partial\theta = 0,$$

is therefore consistent for θ. The robust covariance matrix has the form

$$\mathcal{G}(\theta) = \mathcal{A}'(\theta)\mathcal{B}^{-1}(\theta)\mathcal{A}(\theta)$$

where $\mathcal{A}(\theta) = \mathrm{E}(-\partial S(\theta)/d\theta')$ and $\mathcal{B}(\theta) = \mathrm{E}(S'(\theta)S(\theta))$ are $p \times p$ matrices and $\mathcal{G}(\theta)$ is the Godambe information matrix (Godambe, 1960).

4.2 Likelihood and Composite Likelihood Formulation

4.2.1 Notation and Model Specification

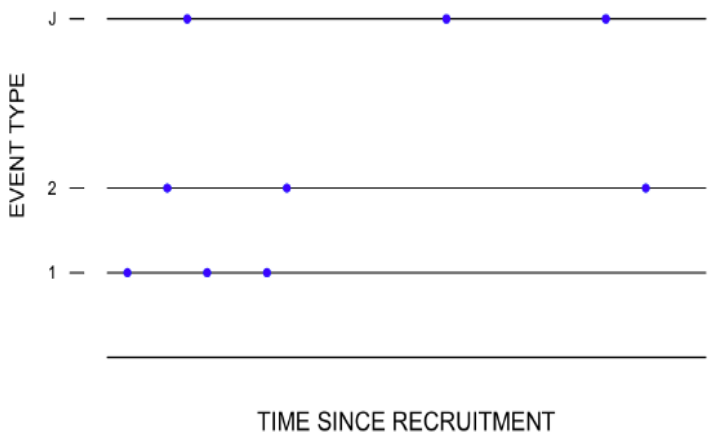


Figure 4.1: Timeline diagrams for J different recurrent event processes and a common censoring time

Suppose individuals are at risk of J types of events and let T_{ijk} denote the time of the kth occurrence type j event for an individual with label i. We let

$$dN_{ij}(s) = I(\text{a type } j \text{ event occurred at time } s \text{ for individual } i)$$

and let $N_{ij}(t) = \int_0^t dN_{ij}(s)$ record the cumulative number of type j events experienced by individual i over $(0, t]$; the corresponding counting process is represented as $\{N_{ij}(u), 0 < u\}$. To consider all J events simultaneously we let $dN_i(s) = (dN_{i1}(s), \ldots, dN_{ij}(s))'$. A $p \times 1$

vector of fixed covariates, possibly unique for type j events, is denoted by x_{ij} and we let $\boldsymbol{x}_i = (x'_{i1}, \ldots, x'_{iJ})'$. The history for type j events is $H_{ij}(t) = \{N_{ij}(s), 0 \leq s < t, x_{ij}\}$ and the full history of all types of events is $H_i(t) = \{H_{i1}(t), \ldots, H_{iJ}(t), 0 \leq s < t\}$.

The complete intensity function for a type j event for individual i is

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N_{ij}(t) = 1 | H_i(t))}{\Delta t} = \lambda_{ij}(t | H_i(t)),$$

$j = 1, \ldots, J$, $i = 1, \ldots, n$.

Suppose each individual in a sample of n independent individuals is to be followed over $(0, A]$ where A is an administrative censoring time. To accommodate possible early study withdrawal, assumed to be conditionally independent of the event processes given X_i, we define a random censoring time $C_i^\dagger$ for individual i and let $C_i = \min(C_i^\dagger, A)$ and $Y_i(s) = I(s \leq C_i)$, $i = 1, \ldots, n$. We then let $d\bar{N}_{ij}(s) = Y_i(s) dN_{ij}(s)$, $d\bar{N}_i(s) = (d\bar{N}_{i1}(s), \ldots, d\bar{N}_{iJ}(s))'$, and $\bar{N}_{ij}(t) = \int_0^t d\bar{N}_{ij}(s)$ which is the observed number of type j events. We let $\bar{H}_{ij}(t) = \{\bar{N}_{ij}(s), Y_i(s), 0 \leq s < t, x_{ij}\}$ and define the observed history $\bar{H}_i(t) = \{\bar{H}_{i1}(t), \ldots, \bar{H}_{iJ}(t), 0 \leq s < t\}$.

In medical research, it is sometimes not possible to fully explain variation between individuals simply by the incorporation of available covariates. To account for variation in risk between individuals we consider a mixed Poisson model in which we introduce random effects; this framework also allows for a dependence between event counts over disjoint intervals (Lawless, 1987; Klein, 1992). We let U_{ij} be a random effect for type j events for individual i. Under a mixed Poisson model subject to independent right censoring, the conditional intensity for observed type j events given the random effect $U_{ij} = u_{ij}$ is

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{ij}(t) = 1 | \bar{H}_i(t), u_{ij})}{\Delta t} = u_{ij} Y_i(t) \lambda_{j0}(t; \alpha_j) \exp(x'_{ij} \beta_j) \tag{4.2.1}$$

where $\lambda_{j0}(t; \alpha_j)$ denotes the baseline event rate function and the covariates have a multiplicative effect. We let $\theta_j = (\alpha'_j, \beta'_j)'$ and $\theta = (\theta'_1, \ldots, \theta'_J)'$.

We let $G_j(u_{ij}; \sigma_j)$ be the c.d.f. for U_{ij} which are i.i.d. for all individuals, and consider the special case in which U_{ij} are log-normal random with $E(U_{ij}) = 1$ and $\text{Var}(\log(U_{ij})) = \sigma_j^2$.

To take into account the dependence between different types of events we consider the joint density of $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iJ})'$ obtained using a Gaussian copula model (Nelsen, 2006). Specifically we let

$$dG(\boldsymbol{u}_i; \sigma, \rho) = \prod_{j=1}^{J} dG_j(u_{ij}; \sigma_j) c(G_1(u_{i1}; \sigma_1), \ldots, G_J(u_{iJ}; \sigma_J); \rho)$$

where $\sigma = (\sigma_1, \ldots, \sigma_J)'$, $G_j(u_{ij}; \sigma_j)$ is the marginal cumulative distribution function of U_{ij}, $j = 1, \ldots, J$, and $c(G_1(u_{i1}; \sigma_1), \ldots, G_J(u_{iJ}; \sigma_J); \rho)$ can be written as

$$c(G_1(u_{i1}; \sigma_1), \ldots, G_J(u_{iJ}; \sigma_J); \rho) = \frac{1}{\sqrt{\det R}} \exp\left(-\tfrac{1}{2} \begin{pmatrix} \Phi^{-1}(G_1(u_{i1}; \sigma_1)) \\ \vdots \\ \Phi^{-1}(G_J(u_{iJ}; \sigma_J)) \end{pmatrix}' (R^{-1} - I) \begin{pmatrix} \Phi^{-1}(G_1(u_{i1}; \sigma_1)) \\ \vdots \\ \Phi^{-1}(G_J(u_{iJ}; \sigma_J)) \end{pmatrix} \right)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal and R is a correlation matrix with (j, k) component ρ_{jk} explaining the association between $G_j(U_{ij}; \sigma_j)$ and $G_k(U_{ik}; \sigma_k)$; we let $\rho = (\rho_{12}, \ldots, \rho_{(J-1)J})$. We let $\psi_j = (\theta_j', \sigma_j)'$ be the marginal parameters, and $\psi = (\psi_1', \ldots, \psi_J')$, and let $\phi = (\sigma', \rho')'$, and overall parameters $\Omega = (\psi', \rho')'$. The marginal likelihood for n independent individual processes is written as

$$L(\Omega) = \prod_{i=1}^{n} \int_0^\infty \cdots \int_0^\infty \left\{ \prod_{j=1}^{J} L_{ij}(\theta_j | u_{ij}) \right\} dG(\boldsymbol{u}_i; \phi) \tag{4.2.2}$$

where the conditional likelihood $L_{ij}(\theta_j | u_{ij})$ is given by

$$L_{ij}(\theta_j | u_{ij}) = \prod_{k=1}^{N_{ij}(C_i)} u_{ij} \lambda_{j0}(t_{ijk}; \alpha_j) \exp(x_{ij}' \beta_j) \exp\left(- \int_0^\infty u_{ij} Y_i(v) \lambda_{j0}(v; \alpha_j) \exp(x_{ij}' \beta_j) dv \right).$$

4.3 Estimation Based on Composite Likelihood

4.3.1 Composite Likelihood Construction

In this setting, when the dimension of J increases, the inference of this model is computationally intractable as there is no closed-form for the full likelihood. We adopt composite likelihood methods to resolve the computational difficulty in estimation. Thus instead of maximizing the full log-likelihood (4.2.2), a pairwise log-likelihood is used for inference. In this case, we consider each pair of events together to determine the composite pairwise likelihood as

$$CL^2(\Omega) = \prod_{i=1}^{n} \prod_{(j,k)\in\mathcal{M}} L_{ijk}(\eta_{jk})^{w_{jk}} \tag{4.3.1}$$

where

$$L_{ijk}(\eta_{jk}) = \int_0^\infty \int_0^\infty L_{ij}(\theta_j|u_{ij}) L_{ik}(\theta_k|u_{ik}) dG_{jk}(u_{ij}, u_{ik}; \phi_{jk}),$$

$$dG_{jk}(u_{ij}, u_{ik}; \phi_{jk}) = dG_j(u_{ij}; \sigma_j) dG_k(u_{ik}; \sigma_k) c(G_j(u_{ij}; \sigma_1), G_k(u_{ik}; \sigma_k); \rho_{jk}),$$

$\eta_{jk} = (\psi_j, \psi_k, \rho_{jk})'$, $\phi_{jk} = (\sigma_j, \sigma_k, \rho_{jk})'$ and $\mathcal{M}$ is the collection of $J(J-1)$ pairs of (j,k) of event types. We note that $w_{jk} = 1/(J-1)$ is chosen to make a single effective contribution of each type of event for each individual to the composite likelihood.

4.3.2 A Semiparametric EM Algorithm for Estimation with Pairwise Likelihood

When margins are specified semiparametrically even solving (4.3.1) directly is difficult due to the high dimension of the parameters. Here we adopt an expectation-maximization algorithm in which we treat random effects as missing data and the data on the event process as observed (Dempster *and others*, 1977). We specify, in this case, $\lambda_{j0}(t_{jk}) = d\Lambda_{j0}(t_{jk})$ as an unspecified function and we estimate $d\Lambda_{j0}(t_{jk}) \equiv \alpha_{jk}$ and let $\theta_j = (d\Lambda'_{j0}, \beta'_j)'$. Given random effects we decompose the complete pairwise composite log-likelihood in (4.3.1) into

the following parts as

$$\log CL^2(\Omega) = \sum_{i=1}^{n} \sum_{(j,k)\in\mathcal{M}} w_{jk}\big(\log L_{ij}(\theta_j|u_{ij}) + \log L_{ik}(\theta_k|u_{ik}) + \log dG_{jk}(u_{ij}, u_{ik}; \phi_{jk})\big) \tag{4.3.2}$$

In the E-step, we take the conditional expectation of the complete pairwise composite log-likelihood given the corresponding paired observed data. Here we define $H_{i,jk}(t) = \{H_{ij}(s), H_{ik}(s), 0 < s < t\}'$ as the history of j and k types of events for an individual i. Then

$$\begin{aligned}\mathrm{E}[\log CL^2(\Omega)|H_{i,jk}(C_i); \widehat{\Omega}^{(r-1)}] = \sum_{i=1}^{n} \sum_{(j,k)\in\mathcal{M}} w_{jk}\big(\mathrm{E}[\log L_{ij}(\theta_j|u_{ij})|H_{i,jk}(C_i); \widehat{\Omega}^{(r-1)}] \\ + \mathrm{E}[\log L_{ik}(\theta_k|u_{ik})|H_{i,jk}(C_i); \widehat{\Omega}^{(r-1)}] + \mathrm{E}[\log dG_{jk}(u_{ij}, u_{ik}; \phi_{jk})\big)|H_{i,jk}(C_i); \widehat{\Omega}^{(r-1)}]\big)\end{aligned} \tag{4.3.3}$$

where $\widehat{\Omega}^{(r-1)}$ is an estimate of Ω at the $(r-1)$st iteration. The estimating equation in the M-step at the rth iteration (Klein, 1992) is

$$U_j^{(r-1)}(\beta_j) = \sum_{i=1}^{n} \int_0^{\infty} \sum_{k=1,j\neq k}^{J} w_{jk} \bar{Y}_i(s) W_{ij}^{(r-1)}(s; \beta_j) dN_{ij}(s) \tag{4.3.4}$$

where

$$W_{ij}^{(r-1)}(s; \beta_j) = \left(x_{ij} - \frac{R_j^{(1,r-1)}(s; \beta_j)}{R_j^{(0,r-1)}(s; \beta_j)} \right), \tag{4.3.5}$$

and we set

$$R_j^{(h,r)}(s; \beta_j) = \sum_{i=1}^{n} \bar{Y}_i(s) \left[\sum_{k=1,j\neq k}^{J} w_{jk} E[U_{ij}|H_{i,jk}(C_i), \widehat{\Omega}^{(r)}] \right] \exp(x_{ij}'\beta_j) x_{ij}^h. \tag{4.3.6}$$

The calculation of $E[U_{ij}|H_{i,jk}(C_i), \widehat{\Omega}^{(r)}]$ is given as

$$E[U_{ij}|H_{i,jk}(C_i); \widehat{\Omega}^{(r)}] = \frac{\int_0^\infty \int_0^\infty u_{ij} P(H_{i,jk}(C_i)|u_{ij}, u_{ik}, x_{ij}, x_{ik}; \widehat{\theta}_j^{(r)}, \widehat{\theta}_k^{(r)}) dG_{jk}(u_{ij}, u_{ik}; \widehat{\phi}_{jk}^{(r)})}{\int_0^\infty \int_0^\infty P(H_{i,jk}(C_i)|u_{ij}, u_{ik}, x_{ij}, x_{ik}; \widehat{\theta}_j^{(r)}, \widehat{\theta}_k^{(r)}) dG_{jk}(u_{ij}, u_{ik}; \widehat{\phi}_{jk}^{(r)})} \tag{4.3.7}$$

where $P(H_{i,jk}(C_i)|u_{ij}, u_{ik}, x_{ij}, x_{ik}; \theta_j, \theta_k) = L_{ij}(\theta_j|u_{ij}) L_{ik}(\theta_k|u_{ik})$ and $\widehat{\theta}_j^{(r)}, \widehat{\theta}_k^{(r)}$ and $\widehat{\phi}_{jk}^{(r)}$ are the estimates of θ_j, θ_k, and ϕ_{jk} at the r-th iteration, respectively. Let $\widehat{\beta}_j^{(r)}$ denote the solution to $U_j^{(r-1)}(\beta_j) = 0$ in (4.3.4). The cumulative baseline rates are then estimated using the Breslow formula as

$$\widehat{\Lambda}_{j0}^{(r)}(s) = \sum_{i=1}^{n} \int_0^\infty \sum_{k=1, j\neq k}^{J} w_{jk} \bar{Y}_i(s) dN_{ij}(s) / R_j^{(0,r-1)}(s; \widehat{\beta}_j^{(r)}) \tag{4.3.8}$$

The maximization of (4.3.3) in semi-parametric setting can be easily carried out using the `coxreg` function in `R` with $\log((\mathrm{E}[U_{ij}|H_{i,jk}(C_i)] + \mathrm{E}[U_{ik}|H_{i,jk}(C_i)])/2)$ treated as an offset term. The variance of the random effects and the dependence parameter are estimated by maximizing

$$\sum_{i=1}^{n} \sum_{(j,k)\in\mathcal{M}} w_{jk} E[\log dG_{jk}(u_{ij}, u_{ik}; \phi_{jk})|H_{i,jk}(C_i), \widehat{\Omega}^{(r-1)}] \tag{4.3.9}$$

using the standard optimization software such as the `optim` function in `R`. The E-step and M-step are repeated iteratively until the following stopping rule is satisfied;

$$\max(|\widehat{\Omega}^{(r+1)} - \widehat{\Omega}^{(r)}|) \leq 10^{-4}.$$

The variances of parameter estimates are obtained by the Godambe information matrix (Godambe, 1960) which is written here as

$$\widehat{G}(\Omega) = \widehat{A}(\Omega)' \widehat{B}(\Omega)^{-1} \widehat{A}(\Omega) \tag{4.3.10}$$

where the estimates of A takes the form using Louis' method (Louis, 1982)

$$\widehat{A} = \sum_{i=1}^{n} \sum_{(j,k)\in\mathcal{M}} w_{jk} \left\{ -E\left[\frac{\partial^2 \log CL^2_{i,jk}}{\partial\Omega\partial\Omega'} \middle| H_{i,jk}(C_i) \right] \Big|_{\Omega=\widehat{\Omega}} - VAR\left[\frac{\partial \log CL^2_{i,jk}}{\partial\Omega} \middle| H_{i,jk}(C_i) \right] \Big|_{\Omega=\widehat{\Omega}} \right\} \tag{4.3.11}$$

and

$$\widehat{B} = \sum_{i=1}^{n} \left\{ \sum_{(j,k)\in\mathcal{M}} w_{jk} E\left[\frac{\partial \log CL^2_{i,jk}}{\partial\Omega} \middle| H_{i,jk}(C_i) \right] \Big|_{\Omega=\widehat{\Omega}} \right\} \left\{ \sum_{(j,k)\in\mathcal{M}} w_{jk} E\left[\frac{\partial \log CL^2_{i,jk}}{\partial\Omega} \middle| H_{i,jk}(C_i) \right] \Big|_{\Omega=\widehat{\Omega}} \right\}' \tag{4.3.12}$$

where $\log CL^2_{i,jk} = \log L_{ij}(\theta_j|u_{ij}) + \log L_{ik}(\theta_k|u_{ik}) + \log dG_{jk}(u_{ij}, u_{ik}; \phi_{jk})$. The details of the variance estimation are presented in the Appendix 4.A. The calculation of conditional expectation in (4.3.7), (4.3.9), (4.3.11), and (4.3.12) requires to use numerical integration. To facilitate shared-memory multi-processor, we implement OpenMP (Open Multi-Processing) interface in `C++` to obtain numerical integration by Gaussian-Quadrature with 20 nodes for each dimension.

4.3.3 Two-stage Semiparametric Estimation with Pairwise Likelihood

The implementation of a two-stage estimation procedure can ease computation (Zhao and Joe, 2005; Andersen, 2004). In the first stage, the parameters for each type of event, $\psi_j = (\theta_j, \sigma_j)'$, are estimated under a working independence assumption. We also use the expectation-maximization algorithm treating a random effect for each type as missing data to obtain $\widehat{\psi}_j$. Given a random effect u_{ij}, the complete likelihood function for each type is $L_{ij}(\theta_j|u_{ij})$. In the E-step, we obtain the conditional expectation given the observed data for each type

$$\mathrm{E}[\log L_{ij}(\theta_j|u_{ij})|H_{ij}(C_i), \widehat{\psi}^{(r)}] \tag{4.3.13}$$

at the r-th iteration. In the M-step, the estimating equation (4.3.4) changes to

$$U_j^{(r-1)}(\beta_j) = \sum_{i=1}^{n} \int_0^{\infty} \bar{Y}_i(s) W_{ij}^{(r-1)}(s;\beta_j) dN_{ij}(s)$$

where $R_j^{(h,r)}(s;\beta_j)$ in (4.3.6) becomes

$$R_j^{(h,r)}(s;\beta_j) = \sum_{i=1}^{n} \bar{Y}_i(s) E[U_{ij}|H_{ij}(C_i), \widehat{\psi}_j^{(r)}] \exp(x_{ij}'\beta_j) x_{ij}^h. \tag{4.3.14}$$

We calculate $E[U_{ij}|H_{ij}(C_i); \widehat{\psi}_j^{(r)}]$ as

$$E[U_{ij}|H_{ij}(C_i); \widehat{\psi}_j^{(r)}] = \frac{\int_0^{\infty} u_{ij} P(H_{ij}(C_i)|u_{ij}, x_{ij}; \widehat{\theta}_j^{(r)}) dG_j(u_{ij}; \widehat{\sigma}_j^{(r)})}{\int_0^{\infty} P(H_{ij}(C_i)|u_{ij}, x_{ij}; \widehat{\theta}_j^{(r)}) dG_j(u_{ij}; \widehat{\sigma}_j^{(r)})}$$

where $P(H_{ij}(C_i)|u_{ij}, x_{ij}; \theta) = L_{ij}(\theta_j|u_{ij})$. The estimated cumulative baseline rates are then obtained as

$$\widehat{\Lambda}_{j0}^{(r)}(s) = \sum_{i=1}^{n} \int_0^{\infty} \bar{Y}_i(s) dN_{ij}(s) / R_j^{(0,r-1)}(s; \widehat{\beta}_j^{(r)}).$$

The procedure is iterated until convergence. In the second stage, we solve the composite score function from the pairwise likelihood with respect to ρ plugging the estimates $\widehat{\psi}$ from the stage 1. Again, we implement the expectation-maximization algorithm with random effects treated as missing data in which we obtain the dependence parameter by maximizing the following estimating equation given the the estimates of marginal parameters

$$\sum_{i=1}^{n} \sum_{(j,k)\in\mathcal{M}} w_{jk} E[\log dG_{jk}(u_{ij}, u_{ik}; \sigma_j, \sigma_k, \rho_{jk})|H_{i,jk}(C_i), \widehat{\rho}^{(r-1)}, \widehat{\psi}_j, \widehat{\psi}_k].$$

Zhao and Joe (2005) commented that two-stage estimation in composite likelihood is recommended with a weak dependence. In a strong dependence case, the simultaneous estimation method gives better estimates. The variance estimates in two-stage estimation are present in Appendix 4.A. The code is available from the author upon request.

4.4 Simulation Studies

Simulation studies were conducted to evaluate the performance of estimators from the joint models introduced in Section 4.2. We consider three different types of infections (J=3). For a randomized treatment, we let x_i the indicator of treatment where $x_i = 1$ having treatment otherwise $x_i = 0$ where $P(X_i = 1) = 0.5$. We generate the data over the interval $(0, 1]$ with an independent random censoring C_i. We assume that C_i follows an exponential distribution with rate $-\log(0.9)$ indicating 10% censoring. The marginal rate functions are of the form $\alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha})$. We set $(\lambda_1, \alpha_1) = (1, 1), (\lambda_2, \alpha_2) = (1.5, 1.25), (\lambda_3, \alpha_3) = (2, 1.25)$ and the coefficients $\beta_1 = \beta_2 = \beta_3 = \log(0.8)$. We consider the Gaussian copula with log-normal margins for random effects where $\mathrm{E}(U_{ij}) = 1$ and $\mathrm{Var}(\log(U_{ij})) = \sigma_j^2$. We set the frailty parameters as $\sigma_1 = 0.4, \sigma_2 = 0.4, \sigma_3 = 0.4$, and the association parameters as $(\rho_{12}, \rho_{13}, \rho_{23}) = (0.25, 0.25, 0.25)$, and $(\rho_{12}, \rho_{13}, \rho_{23}) = (-0.25, -0.25, 0.25)$. We generated 500 samples of 1000 individuals. For each data set, parametric and semi-parametric analyses are carried out based on the pairwise likelihood likelihoods in Section 4.3.2, and two-stage estimation based on the pairwise likelihood in Section 4.3.3. For models with parametric baseline rate functions such as Weibull model, the marginal parameters $\theta_j = (\lambda_j, \alpha_j, \beta_j)'$ can be obtained by maximizing (4.3.3) or (4.3.13) where $\lambda_{j0}(s) = \alpha_j\lambda_j^{\alpha_j}s^{\alpha_j-1}$ using standard optimization software such as the `optim` function in `R`. The empirical bias (EBIAS), average asymptotic (large sample) standard error (ASE), empirical standard error (ESE) and empirical coverage probability (ECP) are evaluated for all parameter estimates and reported in Table 4.1 and 4.2.

The empirical biases are very small for all estimates of parameters and empirical standard errors and average estimated standard errors are in good agreement. The empirical coverage probability are close to the nominal confidence level of 95%. Comparing parametric and semi-parametric model, we find that all estimates and standard errors are very close between two models. However, to protect from the model misspecification, semi-parametric model is recommended although the computation is intense compared to the parametric model. The standard errors of estimators of marginal parameters are very close between the pairwise simultaneous and two-stage model whereas there is some efficiency gain in dependence parameters with the simultaneous model compared to the two-stage

Table 4.1: Frequency properties of estimators obtained by fitting a Weibull-model using the pairwise likelihood and two-stage estimation based on the pairwise likelihood with the sample size 1000 and $nsim = 500$; $(\rho_{12}, \rho_{13}, \rho_{23}) = (0.25, 0.25, 0.25)$

		Pairwise Likelihood				Two-stage Pairwise Likelihood			
TYPE	PARAM	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
				Weibull Model					
1	λ_1	0.000	0.050	0.050	0.952	-0.001	0.050	0.050	0.950
	α_1	0.000	0.035	0.034	0.944	0.001	0.035	0.034	0.942
	β_1	-0.003	0.072	0.074	0.958	-0.002	0.072	0.074	0.958
	σ_1	-0.002	0.064	0.069	0.962	-0.003	0.064	0.070	0.968
2	λ_2	-0.006	0.062	0.063	0.944	-0.007	0.061	0.063	0.950
	α_2	0.000	0.035	0.035	0.948	0.001	0.036	0.035	0.948
	β_2	0.003	0.060	0.062	0.958	0.004	0.060	0.063	0.958
	σ_2	-0.003	0.048	0.051	0.970	-0.002	0.048	0.051	0.974
3	λ_3	-0.007	0.078	0.075	0.934	-0.006	0.077	0.075	0.938
	α_3	0.001	0.031	0.030	0.944	0.001	0.031	0.030	0.942
	β_3	0.005	0.054	0.056	0.956	0.004	0.053	0.056	0.962
	σ_3	-0.003	0.040	0.041	0.960	-0.003	0.040	0.041	0.966
Copula	ρ_{12}	-0.012	0.213	0.234	0.966	-0.008	0.220	0.235	0.960
	ρ_{13}	0.003	0.189	0.207	0.962	0.009	0.194	0.208	0.960
	ρ_{23}	0.007	0.168	0.172	0.956	0.010	0.168	0.172	0.958
				Semiparametric Model					
1	$\Lambda_1(1)$	0.000	0.051	0.050	0.934	-0.001	0.051	0.050	0.934
	β_1	-0.005	0.073	0.074	0.946	-0.006	0.073	0.074	0.946
	σ_1	-0.006	0.062	0.069	0.964	-0.007	0.062	0.070	0.972
2	$\Lambda_2(1)$	0.003	0.061	0.063	0.952	0.003	0.061	0.063	0.950
	β_2	-0.003	0.061	0.062	0.948	-0.003	0.061	0.062	0.948
	σ_2	-0.007	0.052	0.051	0.950	-0.005	0.052	0.051	0.954
3	$\Lambda_3(1)$	0.002	0.082	0.075	0.914	0.002	0.082	0.075	0.912
	β_3	-0.004	0.055	0.056	0.954	-0.004	0.055	0.056	0.956
	σ_3	-0.005	0.039	0.041	0.974	-0.005	0.039	0.041	0.970
Copula	ρ_{12}	-0.010	0.221	0.239	0.964	-0.005	0.226	0.240	0.960
	ρ_{13}	0.009	0.193	0.210	0.966	0.013	0.197	0.212	0.968
	ρ_{23}	0.016	0.171	0.175	0.964	0.017	0.174	0.175	0.958

Table 4.2: Frequency properties of estimators obtained by fitting a Weibull-model using the pairwise likelihood and two-stage estimation based on the pairwise likelihood with the sample size 1000 and $nsim = 500$; $(\rho_{12}, \rho_{13}, \rho_{23}) = (-0.25, -0.25, 0.25)$

		Pairwise Likelihood				Two-stage Pairwise Likelihood			
TYPE	PARAM	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
				Weibull Model					
1	λ_1	-0.001	0.053	0.049	0.940	-0.001	0.053	0.049	0.940
	α_1	0.002	0.035	0.034	0.950	0.002	0.034	0.034	0.950
	β_1	-0.005	0.073	0.074	0.954	-0.005	0.073	0.074	0.954
	σ_1	-0.009	0.061	0.069	0.966	-0.011	0.062	0.071	0.972
2	λ_2	0.002	0.063	0.063	0.952	0.002	0.063	0.063	0.954
	α_2	0.001	0.033	0.035	0.952	0.001	0.034	0.035	0.952
	β_2	-0.003	0.062	0.062	0.938	-0.003	0.062	0.062	0.938
	σ_2	-0.006	0.050	0.050	0.958	-0.006	0.051	0.051	0.962
3	λ_3	0.000	0.078	0.075	0.938	0.000	0.078	0.075	0.936
	α_3	0.002	0.029	0.030	0.946	0.001	0.029	0.030	0.946
	β_3	-0.002	0.056	0.056	0.948	-0.002	0.056	0.056	0.950
	σ_3	-0.005	0.040	0.041	0.942	-0.006	0.040	0.041	0.946
Copula	ρ_{12}	-0.011	0.214	0.244	0.980	-0.018	0.219	0.248	0.980
	ρ_{13}	0.007	0.203	0.216	0.964	0.002	0.209	0.232	0.964
	ρ_{23}	-0.011	0.169	0.174	0.958	-0.010	0.170	0.174	0.956
				Semiparametric Model					
Type1	$\Lambda_1(1)$	-0.001	0.053	0.049	0.940	-0.001	0.053	0.049	0.938
	β_1	-0.005	0.073	0.074	0.956	0.002	0.073	0.074	0.954
	σ_1	-0.009	0.062	0.069	0.960	-0.011	0.062	0.071	0.972
Type2	$\Lambda_2(1)$	0.002	0.063	0.065	0.956	0.002	0.063	0.063	0.952
	β_2	-0.004	0.062	0.062	0.936	-0.003	0.062	0.062	0.936
	σ_2	-0.006	0.050	0.053	0.952	-0.006	0.050	0.051	0.962
Type3	$\Lambda_3(1)$	0.001	0.078	0.077	0.936	0.000	0.078	0.075	0.936
	β_3	-0.002	0.056	0.056	0.950	-0.002	0.056	0.056	0.948
	σ_3	-0.005	0.041	0.042	0.938	-0.005	0.040	0.041	0.946
Copula	ρ_{12}	-0.012	0.213	0.245	0.970	-0.018	0.219	0.248	0.980
	ρ_{13}	0.008	0.203	0.217	0.936	0.002	0.209	0.230	0.966
	ρ_{23}	-0.011	0.169	0.174	0.936	-0.010	0.170	0.174	0.956

model. In this simulation studies, we only consider three types of events in which pair-wise likelihood approach is feasible. However, if more than three events are analyzed two-stage methods are plausible.

4.5 Recurrent Infections in a Pediatric Trial of Iron Supplementation

Lemaire *and others* (2011) conducted a randomized clinical trial of 268 Bangladeshi malnourished children, aged 12-24 month in which children were randomly given iron-containing micro-nutrient powder (iron MNP) or a placebo powder as mentioned in Section 1.3.3. Given this data, interest lies on how to examine the effect of treatment on the incidence of multiple diseases which are caused by malnutrition and iron deficiency. An analysis based on a composite score is feasible (Lemaire *and others*, 2011), however it leads to losing information on distinct episodes. Often interest lies in a treatment effect on specific disease, however, since the lack of nutrition directly or indirectly affects the immune system, the different types of infections in consequence of malnutrition may be associated. Therefore somewhat related diseases should be considered together.

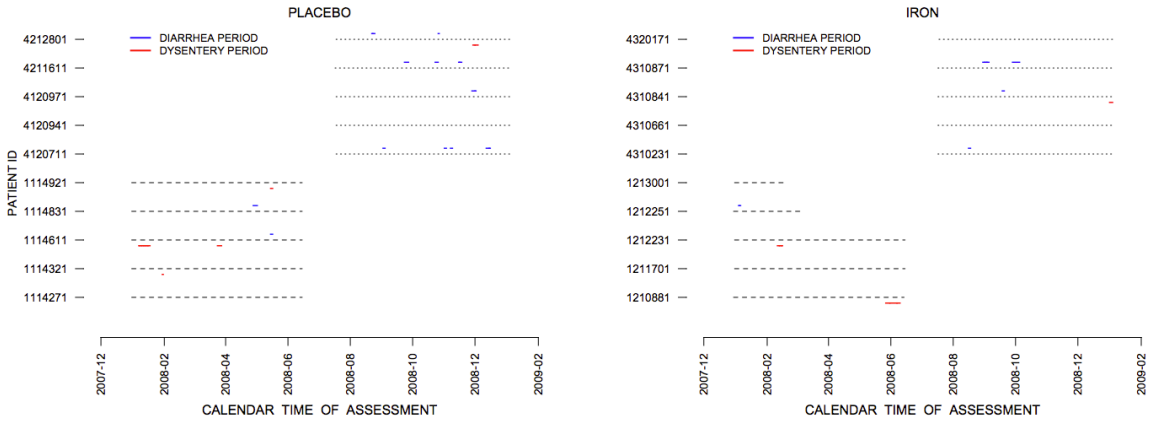


Figure 4.2: Diarrhea, dysentery event plots for phase 1 and phase 2 showing the onset and the duration of episodes

Table 4.3: Joint analysis of three types of infections based on semiparametric model; diarrhea, dysentery and cough with covariates iron and phase

Semiparametric		Pairwise Likelihood			Two-stage PW Likelihood		
TYPE	Covariate	EST	SE	p	EST	SE	p
Coefficients							
Diarrhea	Iron	-0.265	0.140	0.058	-0.265	0.140	0.058
	Phase	0.310	0.138	0.025	0.310	0.138	0.025
Dysentery	Iron	0.200	0.244	0.412	0.202	0.244	0.407
	Phase	0.562	0.248	0.023	0.564	0.246	0.022
Cough	Iron	-0.163	0.108	0.131	-0.163	0.108	0.131
	Phase	0.380	0.109	< 0.001	0.380	0.109	< 0.001
Random effects							
$(\sigma_1, \sigma_2, \sigma_3)$		(0.405,	0.530,	0.314)	(0.372,	0.526,	0.313)
		(0.094)	(0.117)	(0.092)	(0.120)	(0.218)	(0.089)
Dependence Parameters							
$(\rho_{12}, \rho_{13}, \rho_{23})$		0.906	-0.250	-0.379	0.885	-0.276	-0.379
		(0.211)	(0.492)	(0.617)	(0.196)	(0.529)	(0.609)

Figure 4.2 displays diarrhea and dysentery data for each individual with two-phase where lines represent a period of diseases. The onset of diarrhea and dysentery may be correlated, which should be taken into account for an analysis of this data. In our analysis, we select three types of events; diarrhea, dysentery, and cough. Since Figure 4.2 shows that different phase may influence the occurrence of infections we first consider the iron supplement and iron as covariates. We conduct full likelihood and pairwise likelihood in application since the number of subjects is only 268 so that full likelihood is also feasible.

The results using the proposed methods are summarized in Table 4.3. We observe

the iron supplements do not have significant effects on the occurrence of all three types of events. The iron supplements reduce the occurrence of diarrhea (RR : 0.77; 95% CI: (-0.54, 0.01), $p = 0.058$ with pairwise likelihood) and cough (RR : 0.85; 95% CI: (-0.38, 0.05), $p = 0.131$ with pairwise likelihood) whereas it increases the onset of dysentery (RR : 1.22; 95% CI: (-0.28, 0.68), $p = 0.412$ with pairwise likelihood). We note that the change of phase from 1 to 2 significantly increases the occurrence of all type of events; phase 2 represents the winter period so that viral diarrhea increases onset of diarrhea (RR : 1.36; 95% CI: (0.04, 0.58), $p = 0.025$ with pairwise likelihood) and dysentery (RR : 1.75; 95% CI: (0.08, 1.05), $p = 0.023$ with pairwise likelihood). Also the onset of cough may increases due to seasonal factors (RR : 1.46; 95% CI: (0.16, 0.60), $p < 0.001$ with pairwise likelihood). All three types of infections show heterogeneity where $\hat{\sigma}_1 = 0.41, \hat{\sigma}_2 = 0.53, \hat{\sigma}_3 = 0.31$ with pairwise likelihood. From the estimates of dependence parameters, diarrhea infection have a strong positive association with both diarrhea infections ($\hat{\rho}_{12} = 0.91$ with pairwise likelihood). There are negative associations between gastrointestinal infections (diarrhea and dysentery) and cough ($\hat{\rho}_{13} = -0.26, \hat{\rho}_{23} = -0.38$ with pairwise likelihood). If the dependence parameter is close to 1, it may be better to combine two events since the underlying mechanism of two events may be identical. We note that the estimates of covariate effects based on pairwise likelihood and two-stage pairwise likelihood are almost identical, which indicates that marginal covariate effects are robust to dependence parameter.

Figure 4.3 shows the expected number of each episode under the pairwise likelihood for phase 1 and 2, respectively. The infection of diarrhea and cough for the placebo group have approximately 1 event occurrence for the period 12/2007-06/2008 (phase 1) and the iron supplement group has lower expected number of episodes. The expected number of events of dysentery for the placebo group is approximately 0.2 which has low incidence compared to the onset of diarrhea.

4.6 Discussion

We propose the use of a multivariate random effects distribution to model heterogeneity in the risk of several types of recurrent events based on a mixed Poisson model formulation.

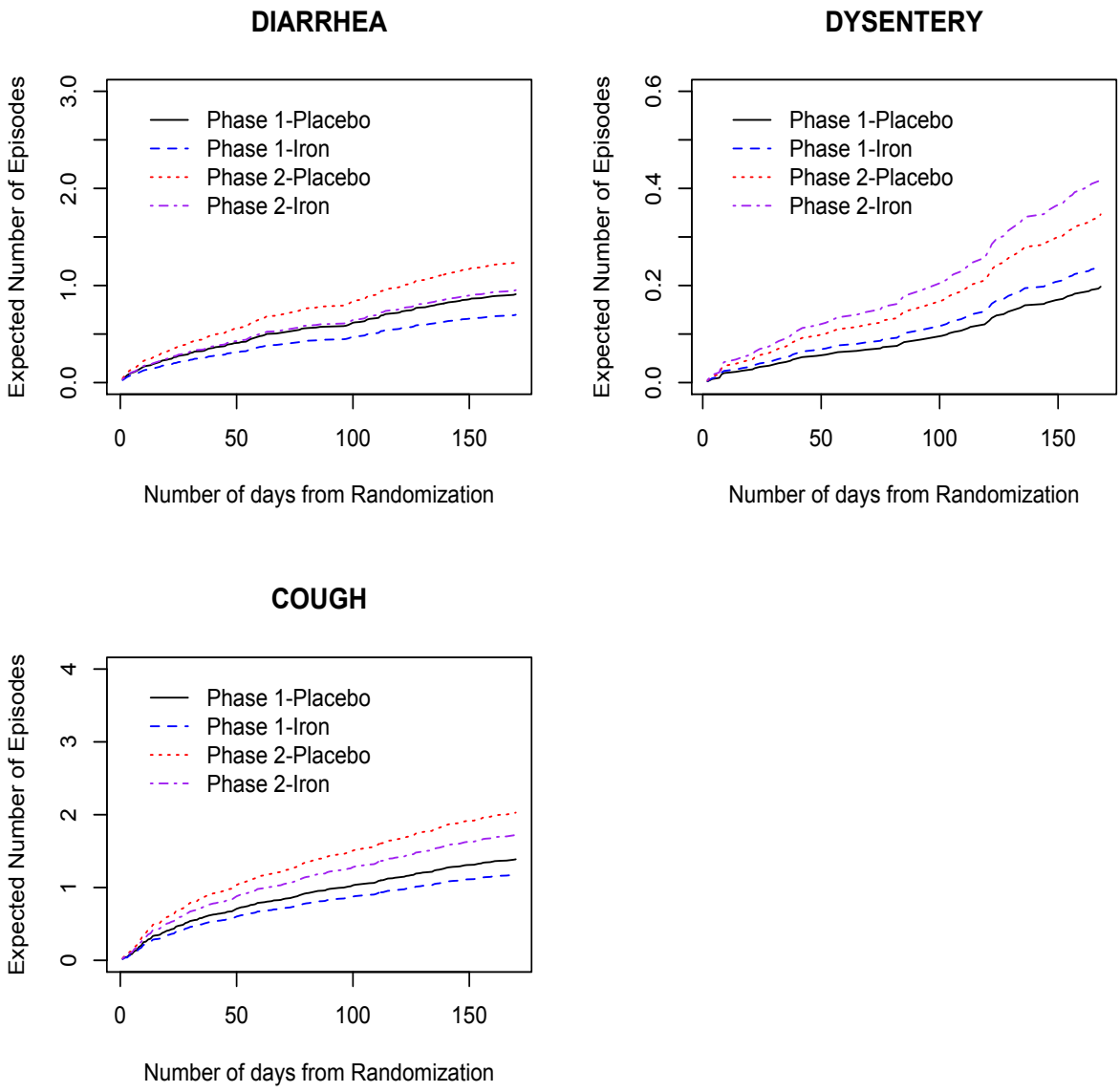


Figure 4.3: Plots of estimated expected number of diarrhea, dysentery, and cough events for placebo and iron MNP group with two phase for the pairwise likelihood analysis using the joint model

The joint distribution of the random effects is constructed via a Gaussian copula model which means that the measures of overdispersion for each type of event are functionally independent and that the dependence structure is quite general. Semiparametric estimation is carried out using a composite likelihood expectation-maximization algorithm; an even more computationally efficient two-stage estimation procedure is also developed which simply uses a working independence assumption at the first stage. Large sample variance estimates are derived for both approaches and are shown to be valid in finite samples in empirical studies. The approach is particularly appealing for use in settings with many different types of events. From intermittent inspection, interval-censoring naturally arises in longitudinal data. We can extend our methods to incorporate this feature of data.

Appendix 4.A: Calculation of the Variance Estimates

1. Expressions for Simultaneous Estimation

To obtain the estimates of covariance matrix in (4.3.10), we need to calculate $\hat{A}$ and $\hat{B}$ in (4.3.11) and (4.3.12), respectively. To be specific,

(i) $\sum_{(j,k)\in\mathcal{M}} w_{jk}\mathrm{E}\left[\frac{\partial^2 \log CL^2 i,jk}{\partial\Omega\partial\Omega'}\middle| H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}} = A_1,$

(ii) $\sum_{(j,k)\in\mathcal{M}} w_{jk}\mathrm{VAR}\left[\frac{\partial \log CL^2 i,jk}{\partial\Omega}\middle| H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}} = A_2,$ and

(iii) $\sum_{(j,k)\in\mathcal{M}} w_{jk}E\left[\frac{\partial \log CL^2_{i,jk}}{\partial\Omega}|H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}} = B_1$

can be obtained as follows. First, we let

$$A_1 = \begin{pmatrix} A_{1,11} & \mathbf{0} \\ \mathbf{0} & A_{1,22} \end{pmatrix} \tag{4.A.1}$$

where

$$A_{1,11} = diag\left(\sum_{(1,k)\in\mathcal{M}} w_{1k}\mathrm{E}\left[\frac{\partial^2 \log L_{i1}(\theta_1|u_{ij})}{\partial\theta_1^2}\middle| H_{i,1k}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}, \ldots, \sum_{(J,k)\in\mathcal{M}} w_{Jk}\mathrm{E}\left[\frac{\partial^2 \log L_{iJ}(\theta_J|u_{iJ})}{\partial\theta_J^2}\middle| H_{i,Jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}\right),$$

and

$$A_{1,22} = \sum_{(j,k)\in\mathcal{M}} w_{jk}\mathrm{E}\left[\frac{\partial^2 \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial\phi\partial\phi'}\middle| H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}.$$

We also let

$$A_2 = \begin{pmatrix} A_{2,11} & A_{2,12} \\ A_{2,21} & A_{2,22} \end{pmatrix} \tag{4.A.2}$$

where

$$A_{2,11} = \begin{cases} w_{jk}\text{Cov}\left[\left(\frac{\partial \log L_{ij}(\theta_j|u_{ij})}{\partial \theta_j}\right), \left(\frac{\partial \log L_{ik}(\theta_k|u_{ik})}{\partial \theta_k}\right) \middle| H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}} & \text{off-diagonal } (j,k) \text{ component} \\ \sum\limits_{(j,k)\in\mathcal{M}} w_{jk}\text{Var}\left[\left(\frac{\partial \log L_{ij}(\theta_j|u_{ij})}{\partial \theta_j}\right) \middle| H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}} & \text{diagonal } (j,j) \text{ component} \end{cases},$$

$$A_{2,12} = A_{2,21}'$$

,

$$A_{2,21} = \sum_{(j,k)\in\mathcal{M}} w_{jk}\text{Cov}\left[\left(\frac{\partial \log dG(u_{ij},u_{ik};\phi_{jk})}{\partial \phi}\right), \left(\frac{\partial \log L_{ij}(\theta_j|u_{ij})L_{ik}(\theta_k|u_{ik})}{\partial \theta}\right) \middle| H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}\Bigg), \text{ and}$$

$$A_{2,22} = \sum_{(j,k)\in\mathcal{M}} w_{jk}\text{Cov}\left[\left(\frac{\partial \log dG(u_{ij},u_{ik};\phi_{jk})}{\partial \phi}\right), \left(\frac{\partial \log dG(u_{ij},u_{ik};\sigma_j,\sigma_k,\rho_{jk})}{\partial \phi}\right) \middle| H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}\Bigg).$$

We also let

$$B_1 = \left[\left(\sum_{(j,k)\in\mathcal{M}} w_{jk}\text{E}\left[\frac{\partial \log L(\theta_j;u_{ij})}{\partial \theta_j}|H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}, j=1,\ldots,J\right), \sum_{(j,k)\in\mathcal{M}} w_{jk}\text{E}\left[\frac{\partial \log dG(u_{ij},u_{ik};\phi_{jk})}{\partial \phi}\middle|H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}\right]'.$$

2. Two-stage Estimation

In two-stage estimation, $A_{1,11}$ and $A_{1,22}$ in (4.A.1) can be written as

$$A_{1,11} = diag\left(\text{E}\left[\frac{\partial^2 \log L_{i1}(\theta_1|u_{ij})}{\partial \theta_1^2}\middle|H_{i1}(C_i)\right]\Big|_{\psi_1=\widehat{\psi}_1}, \ldots, \text{E}\left[\frac{\partial^2 \log L_{iJ}(\theta_J|u_{iJ})}{\partial \theta_J^2}\middle|H_{iJ}(C_i)\right]\Big|_{\psi_J=\widehat{\psi}_J}\right),$$

and

$$A_{1,22} = \begin{pmatrix} diag\left(\text{E}\left[\frac{\partial^2 \log G(u_{ij};\sigma_j)}{\partial \sigma_j^2}|H_{ij}(C_i)\right]\Big|_{\psi_j=\widehat{\psi}_j}; j=1,\ldots,J\right) & \mathbf{0} \\ \sum\limits_{(j,k)\in\mathcal{M}} w_{jk}\text{E}\left[\frac{\partial^2 \log c(u_{ij},u_{ik};\phi_{jk})}{\partial \rho_{jk}\partial \sigma'}|H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}} & diag\left(w_{jk}\text{E}\left[\frac{\partial^2 \log c(u_{ij},u_{ik};\phi_{jk})}{\partial \rho_{jk}^2}|H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}; (j,k)\in\mathcal{M}\right) \end{pmatrix}.$$

Also, in (4.A.2) $A_{2,11}$ becomes

$$A_{2,11} = diag\left(\text{Var}\left[\frac{\partial \log L_{ij}(\theta_j; u_{ij})}{\partial \theta_j}|H_{ij}(C_i)\right]\Big|_{\psi_j=\widehat{\psi}_j}; j = 1, \ldots, J\right),$$

$$A_{2,12} = \left(diag\left(\text{Cov}\left[\tfrac{\partial \log L_{ij}(\theta_j;u_{ij})}{\partial \theta_j}, \tfrac{\partial \log G(u_{ij};\sigma_j)}{\partial \sigma_j}|H_{ij}(C_i)\right]\Big|_{\psi_j=\widehat{\psi}_j}; j = 1, \ldots, J\right) \quad \mathbf{0}\right),$$

$$A_{2,21} = \begin{pmatrix} diag\left(\text{Cov}\left[\tfrac{\partial \log G(u_{ij};\sigma_j)}{\partial \sigma_j}, \tfrac{\partial \log L_{ij}(\theta_j;u_{ij})}{\partial \theta_j}|H_{ij}(C_i)\right]\Big|_{\psi_j=\widehat{\psi}_j}; j = 1, \ldots, J\right) \\ \sum_{(j,k)\in\mathcal{M}} w_{jk}\text{Cov}\left[\left(\tfrac{\partial \log dG(u_{ij},u_{ik};\phi_{jk})}{\partial \phi}\right), \left(\tfrac{\partial \log L_{ij}(\theta_j|u_{ij})L_{ik}(\theta_k|u_{ik})}{\partial \theta}\right)\Big|H_{i,jk}(C_i),\right]\Big|_{\Omega=\widehat{\Omega}} \end{pmatrix},$$

and

$$A_{2,22} = \begin{pmatrix} diag\left(\text{Var}\left[\tfrac{\partial \log G(u_{ij};\sigma_j)}{\partial \sigma_j}|H_{ij}(C_i)\right]\Big|_{\psi_j=\widehat{\psi}_j}, j = 1, \ldots, J\right) & \mathbf{0} \\ \sum_{(j,k)\in\mathcal{M}} w_{jk}\text{Cov}\left[\left(\tfrac{\partial \log dG(u_{ij},u_{ik};\phi_{jk})}{\partial \rho_{jk}}\right), \left(\tfrac{\partial \log dG(u_{ij},u_{ik};\phi_{jk})}{\partial \phi}\right)\Big|H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}} & \end{pmatrix}.$$

In addition, B_1 can be written as

$$B_1 = \Bigg[\left(\text{E}\left[\frac{\partial \log L(\theta_j; u_{ij})}{\partial \theta_j}|H_{ij}(C_i)\right]\Big|_{\psi_1=\widehat{\psi}_1}, j = 1, \ldots, J\right), \left(\text{E}\left[\frac{\partial \log dG(u_{ij};\sigma_j)}{\partial \sigma_j}\Big|H_{ij}(C_i)\right]\Big|_{\psi_J=\widehat{\psi}_J}, j = 1, \ldots, J\right),$$
$$\left(\sum_{(j,k)\in\mathcal{M}} w_{jk}\text{E}\left[\frac{\partial \log c(u_{ij}, u_{ik}; \phi_{jk})}{\partial \rho_{jk}}|H_{i,jk}(C_i)\right]\Big|_{\Omega=\widehat{\Omega}}; (j,k) \in \mathcal{M}\right)\Bigg]'.$$

3. The Conditional Score Vector and the Conditional Information Matrix

The conditional score vector $\partial L_{ij}(\theta_j|u_{ij})/\partial\theta_j$ for $\beta_j, d\Lambda_{j0}(\cdot)$ in the complete pairwise likelihood (4.3.2) are given as

$$\frac{\partial L_{ij}(\theta_j|u_{ij})}{\partial\beta_j} = \sum_{i=1}^{n}\left\{N_{ij}(C_i)x_{ij} - u_{ij}\int_0^\infty \bar{Y}_i(v)x_{ij}\exp(x'_{ij}\beta_j)d\Lambda_{j0}(v)\right\},$$
$$\frac{\partial L_{ij}(\theta_j|u_{ij})}{\partial d\Lambda_{j0}(t_{jk})} = \frac{1}{d\Lambda_{j0}(t_{jk})} - \sum_{i=1}^{n} u_{ij}\bar{Y}_i(t_{jk})\exp(x'_{ij}\beta_j),$$

where t_{jk} is the k^{th} time of type j event occurrence.

The components of the conditional information matrix $-\partial^2 L_{ij}(\theta_j|u_{ij})/\partial\theta_j\partial\theta'_j$ are as follows.

$$-\frac{\partial^2 L_{ij}(\theta_j|u_{ij})}{\partial\beta_j\partial\beta'_j} = \sum_{i=1}^{n} u_{ij}\int_0^\infty \bar{Y}_i(v)x_{ij}x'_{ij}\exp(x'_{ij}\beta_j)d\Lambda_{j0}(v),$$
$$-\frac{\partial^2 L_{ij}(\theta_j|u_{ij})}{\partial\beta_j\partial d\Lambda_{j0}(t_{jk})} = \sum_{i=1}^{n} u_{ij}\bar{Y}_i(t_{jk})x_{ij}\exp(x'_{ij}\beta_j), \text{ and}$$
$$-\frac{\partial^2 L_{ij}(\theta_j|u_{ij})}{\partial\{d\Lambda_{j0}(t_{jk})\}^2} = \frac{1}{d\Lambda_{j0}(t_{jk})^2}.$$

In (4.3.2), $\log dG(u_{ij}, u_{ik}; \phi_{jk})$ is given as

$$\log dG(u_{ij}, u_{ik}; \phi_{jk}) = \sum_{i=1}^{n}\left[\frac{1}{2}\log(u_{ij}) - \frac{\log^2(u_{ij})}{2\sigma_j^2} - \log(\sigma_j) - \frac{\sigma_j^2}{8} + \frac{1}{2}\log(u_{ik}) - \frac{\log^2(u_{ik})}{2\sigma_k^2}\right.$$

$$
\begin{aligned}
&- \log(\sigma_k) - \frac{\sigma_k^2}{8} - \frac{\log(1-\rho_{jk}^2)}{2} - \frac{\rho_{jk}^2(\sigma_j^2+\sigma_k^2)}{8(1-\rho_{jk}^2)} + \frac{\rho_{jk}\sigma_j\sigma_k}{4(1-\rho_{jk}^2)} \\
&- \frac{\rho_{jk}^2}{2(1-\rho_{jk}^2)}\left\{\frac{\log^2(u_{ij})}{\sigma_j^2} + \log(u_{ij}) + \frac{\log^2(u_{ik})}{\sigma_k^2} + \log(u_{ik})\right\} \\
&+ \frac{\rho_{jk}}{1-\rho_{jk}^2}\left\{\frac{\log(u_{ij})\log(u_{ik})}{\sigma_j\sigma_k} + \frac{\sigma_k\log(u_{ij})}{2\sigma_j} + \frac{\sigma_j\log(u_{ik})}{2\sigma_k}\right\}\Bigg].
\end{aligned}
$$

The components of the conditional score vector $\partial \log dG(u_{ij}, u_{ik}; \phi_{jk})/\partial\phi_{jk}$ are given as follows.

$$
\begin{aligned}
\frac{\partial \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial \sigma_j} &= -\frac{n}{\sigma_j} - \frac{n\sigma_j}{4(1-\rho_{jk}^2)} + \frac{n\rho_{jk}\sigma_k}{4(1-\rho_{jk}^2)} + \sum_{i=1}^{n}\frac{\log^2(u_{ij})}{(1-\rho_{jk}^2)\sigma_j^3} \\
&\quad - \sum_{i=1}^{n}\frac{\rho_{jk}\log(u_{ij})\log(u_{ik})}{(1-\rho_{jk}^2)\sigma_j^2\sigma_k} + \frac{\rho_{jk}}{1-\rho_{jk}^2}\sum_{i=1}^{n}\left\{\frac{-\sigma_k\log(u_{ij})}{2\sigma_j^2} + \frac{\log(u_{ik})}{2\sigma_k}\right\}, \\
\frac{\partial \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial \sigma_k} &= -\frac{n}{\sigma_k} - \frac{n\sigma_k}{4(1-\rho_{jk}^2)} + \frac{n\rho_{jk}\sigma_j}{4(1-\rho_{jk}^2)} + \sum_{i=1}^{n}\frac{\log^2(u_{ik})}{(1-\phi_0^2)\sigma_k^3} \\
&\quad - \sum_{i=1}^{n}\frac{\phi_0\log(u_{ij})\log(u_{ik})}{(1-\phi_0^2)\sigma_j\sigma_k^2} + \frac{\rho_{jk}}{1-\rho_{jk}^2}\sum_{i=1}^{n}\left\{\frac{-\sigma_j\log(u_{ik})}{2\sigma_k^2} + \frac{\log(u_{ij})}{2\sigma_j}\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial \rho_{jk}} &= \frac{n\rho_{jk}}{1-\rho_{jk}^2} - \frac{n\rho_{jk}(\sigma_j^2+\sigma_k^2)}{4(1-\rho_{jk}^2)^2} + \frac{n\sigma_j\sigma_k(1+\rho_{jk}^2)}{4(1-\rho_{jk}^2)^2} \\
&\quad - \frac{\rho_{jk}}{(1-\rho_{jk}^2)^2}\sum_{i=1}^{n}\left\{\frac{\log^2(u_{ij})}{\sigma_j^2} + \frac{\log^2(u_{ik})}{\sigma_k^2} + \log(u_{ij}) + \log(u_{ik})\right\} \\
&\quad + \frac{1+\rho_{jk}^2}{(1-\rho_{jk}^2)^2}\sum_{i=1}^{n}\left\{\frac{\log(u_{ij})\log(u_{ik})}{\sigma_j\sigma_k} + \frac{\sigma_k\log(u_{ij})}{2\sigma_j} + \frac{\sigma_j\log(u_{ik})}{2\sigma_k}\right\}.
\end{aligned}
$$

The elements of the conditional information matrix $-\partial^2 \log dG(u_{ij}, u_{ik}; \phi_{jk})/\partial\phi_{jk}\partial\phi'_{jk}$ are as follows.

$$
\begin{aligned}
-\frac{\partial^2 \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial\sigma_j\partial\sigma_j} &= -\frac{n}{\sigma_j^2} + \frac{n}{4(1-\rho_{jk}^2)} + \sum_{i=1}^{n} \frac{3\log^2(u_{ij})}{\sigma_j^4(1-\phi_0^2)} \\
&\quad - \sum_{i=1}^{n} \frac{2\rho_{jk}\log(u_{ij})\log(u_{ik})}{(1-\rho_{jk}^2)\sigma_j^3\sigma_k} - \frac{\rho_{jk}}{1-\rho_{jk}^2}\sum_{i=1}^{n}\frac{\sigma_k\log(u_{ij})}{\sigma_j^3}, \\
-\frac{\partial^2 \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial\sigma_k\partial\sigma_k} &= -\frac{n}{\sigma_k^2} + \frac{n}{4(1-\rho_{jk}^2)} + \sum_{i=1}^{n} \frac{3\log^2(u_{ik})}{\sigma_k^4(1-\rho_{jk}^2)} \\
&\quad - \sum_{i=1}^{n} \frac{2\rho_{jk}\log(u_{ij})\log(u_{ik})}{(1-\rho_{jk}^2)\sigma_j\sigma_k} - \frac{\rho_{jk}}{1-\rho_{jk}^2}\sum_{i=1}^{n}\frac{\sigma_j\log(u_{ik})}{\sigma_k^3}, \\
-\frac{\partial^2 \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial\sigma_j\partial\sigma_k} &= -\frac{n\rho_{jk}}{4(1-\rho_{jk}^2)} - \sum_{i=1}^{n} \frac{\rho_{jk}\log(u_{ij})\log(u_{ik})}{(1-\rho_{jk}^2)\sigma_j^2\sigma_k^2} \\
&\quad + \frac{\rho_{jk}}{1-\rho_{jk}^2}\sum_{i=1}^{n}\left\{\frac{\log(u_{ij})}{2\sigma_j^2} + \frac{\log(u_{ik})}{2\sigma_k^2}\right\}, \\
-\frac{\partial^2 \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial\sigma_j\partial\rho_{jk}} &= \frac{n\rho_{jk}\sigma_j}{2(1-\rho_{jk}^2)^2} - \frac{n\sigma_k(1+\rho_{jk}^2)}{4(1-\rho_{jk}^2)^2} - \sum_{i=1}^{n}\frac{2\rho_{jk}\log^2(u_{ij})}{(1-\rho_{jk})^2\sigma_j^3} \\
&\quad + \frac{(1+\rho_{jk}^2)}{(1-\rho_{jk}^2)^2}\sum_{i=1}^{n}\left\{\frac{\log(u_{ij})\log(u_{ik})}{\sigma_j^2\sigma_k} - \frac{\sigma_k\log(u_{ij})}{\sigma_j^2} + \frac{\log(u_{ik})}{2\sigma_k}\right\},
\end{aligned}
$$

$$
\begin{aligned}
-\frac{\partial^2 \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial \sigma_k \partial \rho_{jk}} &= \frac{n\rho_{jk}\sigma_j}{2(1-\rho_{jk}^2)^2} - \sum_{i=1}^{n} \frac{2\rho_{jk}\log^2(u_{ik})}{(1-\rho_{jk})^2\sigma_k^3} \\
&\quad + \frac{(1+\rho_{jk}^2)}{(1-\rho_{jk}^2)^2} \sum_{i=1}^{n} \left\{ \frac{\log(u_{ij})\log(u_{ik})}{\sigma_j\sigma_k^2} - \frac{\sigma_j \log(u_{ik})}{\sigma_k^2} + \frac{\log(u_{ij})}{2\sigma_k} \right\}, \text{ and} \\
-\frac{\partial \log dG(u_{ij}, u_{ik}; \phi_{jk})}{\partial \rho_{jk} \partial \rho_{jk}} &= -\frac{n(1+\rho_{jk}^2)}{(1-\rho_{jk}^2)^2} + \frac{(\sigma_j^2+\sigma_k^2)(1+3\rho_{jk}^2)}{4(1-\rho_{jk}^2)^3} - \frac{\sigma_j\sigma_k(3\rho_{jk}+\rho_{jk}^3)}{2(1-\rho_{jk}^2)^3} \\
&\quad + \frac{1+3\rho_{jk}^2}{(1-\rho_{jk}^2)^3} \sum_{i=1}^{n} \left\{ \frac{\log^2(u_{ij})}{\sigma_j^2} + \frac{\log^2(u_{ik})}{\sigma_k^2} + \log(u_{ij}) + \log(u_{ik}) \right\} \\
&\quad - \frac{2\rho_{jk}(3+\rho_{jk}^2)}{(1-\rho_{jk}^2)^3} \sum_{i=1}^{n} \left\{ \frac{\log(u_{ij})\log(u_{ik})}{\sigma_j\sigma_k} + \frac{\sigma_k \log(u_{ij})}{2\sigma_j} + \frac{\sigma_j \log(u_{ik})}{2\sigma_k} \right\}.
\end{aligned}
$$

In two-stage estimation, $\log dG(u_{ij}; \sigma_j)$ is given as

$$
\log dG(u_{ij}; \sigma_j) = \sum_{i=1}^{n} \left[\frac{1}{2}\log(u_{ij}) - \frac{\log^2(u_{ij})}{2\sigma_j^2} - \log(\sigma_j) - \frac{\sigma_j^2}{8} \right].
$$

The conditional score function and the conditional information function for σ_j from $\log dG(u_{ij}; \sigma_j)$ are given as

$$
\begin{aligned}
\frac{\partial \log dG(u_{ij}; \sigma_j)}{\partial \sigma_j} &= -\frac{n}{\sigma_j} + \frac{\log^2(u_{ij})}{\sigma_j^3} - \frac{\sigma_j^3}{4}, \text{ and} \\
-\frac{\partial^2 \log dG(u_{ij}; \sigma_j)}{\partial \sigma_j \partial \sigma_j} &= -\frac{n}{\sigma_j^2} + \frac{3\log^2(u_{ij})}{\sigma_j^4} + \frac{3\sigma_j^4}{4}.
\end{aligned}
$$

Chapter 5

The Illness-Death Model for Family Studies

5.1 Introduction

Family studies are conducted to assess the nature and extent of familial aggregation of disease, as well as the effect of genetic factors on disease onset. When present, familial aggregation suggests a shared genetic or environmental basis of disease (Li *and others*, 1998; Liang and Beaty, 2000). For valid inference in such settings, however, it is important to address the sampling scheme by which family members are recruited. Typically an individual with the disease, called the proband, is recruited to the study and provides a detailed disease history including the age of disease onset. The disease onset time for this individual is right-truncated since they had to be affected to be sampled, but their survival time is left-truncated. The family members of the proband, called non-probands, are then selected for the family study. In some settings the proband may report the disease history of their family members, but it may alternatively be acquired through clinical examination conducted by a physician.

A variety of methods for the analysis of multivariate failure time data methods have been developed (Hougaard, 2000). A non-parametric estimate was suggested by Dabrowska

(1988). Wei *and others* (1989) introduced marginal models for different types of failure and developed methods for the robust estimation of standard errors. The marginal apporach for the analysis of clustered failure time data has been developed in general by Lee *and others* (1992) and by Liang *and others* (1993) for family studies. Clayton (1978) suggested use of the cross-ratio as a dependence measure, and Oakes (1989) showed the connection between frailty models and the cross-ratio hazard function. Frailty models have been widely used in the analysis of case-control family studies (Hsu *and others*, 2004; Hsu and Gorfine, 2005) where a frailty variance is interpreted as a measure of dependence in the age of onset within family members. Copula models can alternatively be used, in which case the multivariate joint distribution is formulated in terms of the marginal distributions and a copula function (Joe, 1997; Shih and Louis, 1995). Li *and others* (1998), Shih and Chatterjee (2002), and Chatterjee *and others* (2006) developed the copula models for case-control family studies considering the ascertainment of case-control probands. Zhong and Cook (2016) used copula functions and composite likelihood for the analysis of right-censored and current status family data while addressing complex sampling schemes. Zhong and Cook (2017) developed estimating function methods and considered the implications of different forms of the estimating functions in terms of robustness and efficiency.

The aforementioned methods focus on modeling familial aggregation in disease onset times in the simple framework of time to event data. More recent work has dealt with clustered failure time data in the semi-competing risks setting, where disease onset and disease-free death are considered as competing events. Bandeen-Roche and Liang (2002) suggested a modified conditional hazard ratio to account for the cause of failure based on a frailty model and applied it to a population cohort study of dementia. Shih and Albert (2010) extended the work of Bandeen-Roche and Liang (2002) and considered two types of dependence measures with one to model the dependence in terms of the failure time of paired members and a second to model the association between the failure types given the time; they suggested use of a time-varying piecewise constant dependence measure. To examine sibship association in disease onset, Cheng *and others* (2009) developed nonparametric association analysis using the bivariate cumulative incidence function defined by the cause-specific hazard function to account for the exchangeable clustered competing risks setting. Zhou *and others* (2012) proposed a marginal proportional subdistribution hazard

model in the clustered competing risks setting. Scheike *and others* (2010) and Scheike and Sun (2012) studied a semiparametric additive model and explored a cross-odds ratio-type measure on the probability scale as the association parameters for the Danish twin data; Scheike *and others* (2014) extended the model to accommodate delayed entry and to accommodate genetic and environmental effects.

Multi-state models offer another framework for dependence modeling. Aalen *and others* (1980) applied the Schweder (1970) concept of local dependence to understand the interaction between two life history events by comparing the transition intensities. Hougaard *and others* (1992) and Hougaard (1999) considered dependence modeling in the life times of twins via multistate models under the Markov or semi-Markov assumption.

There has been little work on the use of illness-death models in the setting of family or twin studies. The illness-death model is a useful framework for event history analysis when not only disease incidence but also mortality is considered for better understanding of the life history process (Andersen, 1988). Dependence modeling of correlated illness-death processes is necessary when data are clustered as in family studies. (Cook and Lawless, 2018, Section 6.2) discuss a variety of methods for dependence modeling for clustered or otherwise correlated multistate processes. Some of these apply generally while much of the discussions invlovles progressive process. Jiang and Haneuse (2017) proposed an illness-death model with a non-parametric frailty distribution where the non-terminal event times and terminal event times are correlated.

In this chapter, we develop an illness-death model using the latent variable formulation of the competing risk model for the first event (disease onset or disease-free death). A copula model is used to accommodate clustering in the ages of disease onset within families. Methods are described which account for incomplete data under two types of biased sampling schemes. The use of auxiliary data is highlighted to address identifiability problems and to increase the efficiency. Finally, we show how to account for incomplete genetic data when auxiliary data do not have genotype information available.

The remainder of this contribution is organized as follows. In Section 5.2 we define notation and present the joint model. Two study designs under the biased sampling schemes are then described and the associated likelihood is presented; composite likelihood is pro-

posed for settings where some family sizes are large. The use of auxiliary data is discussed in Section 5.3 to facilitate estimation of transition intensities to the death state, and simulation studies are reported in Section 5.4. In Section 5.5, we extend the proposed methods to incorporate genotype information and present the results of further simulation studies. An application to a family study on the onset of psoriatic arthritis (PsA) from the University of Toronto is given in Section 5.6 and concluding remarks are given in Section 5.7. Remarks on computational methods are given in Appendix 5.A and the method for modeling missing genetic data is described brifely in Appendix 5.B

5.2 Model Formulation

5.2.1 Notation and Model Formulation

We consider a four-state model illness-death model to describe the joint distribution of disease onset and death (Datta *and others*, 2000; Xu *and others*, 2010). We let state 0 represent a healthy state, state 1 represent a diseased state, state 2 represent death post-disease, and state 3 represent disease-free death; see Figure 5.1. Our initial interest lies in modeling the dependence between family members in the age of disease onset. To discuss the joint model in the greatest simplicity, we first consider dependence modeling for individuals labeled j and k in family i, and define variables for individual j without loss of generality. We let X_{ij1} denote the age of disease onset, X_{ij2} the age at death following disease, X_{ij3} the age at disease-free death. Note that this is a latent variable formulation of the semi-competing risks problem for transition out of state 0 in that X_{ij1} may not be observed (or realized) if $X_{ij3} < X_{ij1}$. While unconventional and not without limitations in terms of its connection with observable features, we adopt this formulation here since the association in the age of disease onset is most naturally modeled in terms of the $0 \to 1$ transition times. Finally we let B_{ij} be the calendar time of birth for individual j in family i, $j = 1, 2$, and let $\boldsymbol{B}_i = (B_{ij}, B_{ik})'$ be the vector of calendar times of births for individuals j and k in family i.

It is also helpful to use notation for multistate models and so we let $Z_{ij}(a)$ denote the state occupied at age a and calendar time $B_{ij} + a$ for individual j in family $i = 1, \ldots, n_F$

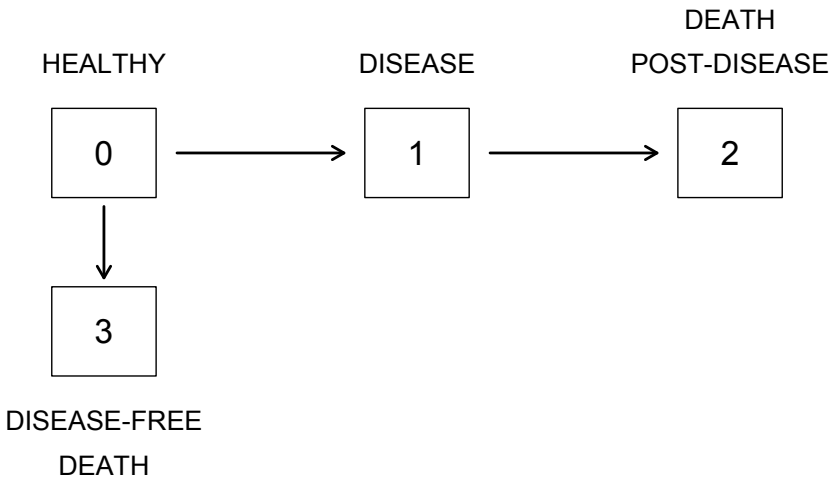


Figure 5.1: A four-state representation of an illness-death model

where n_F is the number of families recruited. We let V_{ij} denote covariates for individual j in family i and $\boldsymbol{V}_i = (V_{ij}, V_{ik})'$. Let $H_{ij}(a) = \{Z_{ij}(s), 0 \leq s < a, B_{ij}, V_{ij}\}$ denote the history for individual j in family i over age $[0, a)$ whose calendar time of birth is B_{ij}. The age- and calendar time-specific marginal transition intensity function from state h into l is defined as

$$\lim_{\Delta a \downarrow 0} \frac{P(Z_{ij}(a + \Delta a^-) = l | Z_{ij}(a^-) = h, H_{ij}(a))}{\Delta a} = \lambda_{ijl}(t, a | H_{ij}(a))$$

with $t = B_{ij} + a$ where $(h, l) \in \{(0, 1), (0, 3), (1, 2)\}$. If the disease process for each family member is Markov given the date of birth and the covariates, we can write

$$\lambda_{ij}(t, a | H_{ij}(a)) = \lambda_l(t, a | b_{ij}, v_{ij}), \qquad l = 1, 2, 3.$$

If we assume that $\lambda_3(t, a|b_{ij}, v_{ij}) = \lambda_2(t, a|b_{ij}, v_{ij})$, the disease is incidental in that it does not change the risk of death, but if $\lambda_3(t, a|b_{ij}, v_{ij}) \neq \lambda_2(t, a|b_{ij}, v_{ij})$, then survival is locally dependent of the disease process (Aalen, 2012); in this case, typically, $\lambda_2(t, a|b_{ij}, v_{ij}) > \lambda_3(t, a|b_{ij}, v_{ij})$. Andersen *and others* (1985) use a Cox model to accommodate proportional

mortality among diseased and disease-free individuals. In the present context, this has the form

$$\lambda_2(t, a|b_{ij}, v_{ij}) = \lambda_3(t, a|b_{ij})\nu_0(a)\exp(v'_{ij}\beta_2), \tag{5.2.1}$$

where $\lambda_3(t, a|b_{ij})$ is the age- and calendar-time specific baseline population mortality rate. The term $\nu_0(a)$ reflects a proportional change in mortality in terms of age. We adopt the model (5.2.1) and let $\lambda_3(t, a|b_{ij}, v_{ij}) = \lambda_3(t, a|b_{ij})$ so that the subject-specific disease-free mortality rate does not depend on the covariates and can be therefore easily estimated based on the population rates. If the disease does not alter the mortality trend in age in the population, we may assume $\nu_0(a) = \nu$.

Here, the transition times X_{ij1} and X_{ij3} are defined to be statistically independent, and we note that we only observe $\min(X_{ij1}, X_{ij3})$ (Kalbfleisch and Prentice, 2011). For transitions from the healthy to diseased state, we assume that the intensity does not depend only on calendar time, in which case we may write $\lambda_1(t, a|b_{ij}, v_{ij}) = \lambda_1(a)\exp(v'_{ij}\beta_1)$. Under the assumption of independent competing risks, we denote the survival functions for the latent disease onset time as $\mathcal{F}(a|V_{ij};\phi_1) = 1-\exp(-\int_0^a \lambda_1(a|V_{ij})da)$, and let $f(a|V_{ij};\phi_1) = -\partial\mathcal{F}(a|V_{ij};\phi_1)/\partial a$ where ϕ_1 indexes the marginal intensity for disease onset. To identify the nature of familial aggregation in the age of disease onset, we construct a joint model for X_{ij1} and X_{ik1} using a copula function (Joe, 1997) in which

$$P(X_{ij1} > a_j, X_{ik1} > a_k; \boldsymbol{V}_i, \varphi) = \mathcal{C}(\mathcal{F}(a_j|V_{ij};\phi_1), \mathcal{F}(a_k|V_{ik};\phi_1); \rho), \tag{5.2.2}$$

where ρ indexes the copula function and $\varphi = (\phi'_1, \rho)'$. We define $\phi = (\phi'_1, \phi'_2)'$ where ϕ_2 indexes the transition intensity from the diseased to death state and $\psi = (\phi', \rho)$. The joint density function can be written as

$$P(X_{ij1} = a_j, X_{ik1} = a_k; \boldsymbol{V}_i, \varphi) = c(\mathcal{F}(a_j|V_{ij};\phi_1), \mathcal{F}(a_k|V_{ik};\phi_1); \rho) f(a_j|V_{ij};\phi_1) f(a_k|V_{ik};\phi_1),$$

where $c(\cdot,\cdot;\rho)$ is the density of the copula. We here consider the Clayton copula which has the form

$$\mathcal{C}(u_1, u_2; \rho) = (u_1^{-\rho} + u_2^{-\rho} - 1)^{-1/\rho}$$

with Kendall's $\tau = \rho/(\rho+2)$. Mesfioui and Quessy (2008) showed that the conditional

Clayton copula has an invariance property; in the present context, if $X_{i1} = (X_{ij1}, X_{ik1}, X_{il1})$ follows a joint distribution with a dependence model governed by the Clayton copula, then the distribution of $X_{ij1}, X_{ik1}|X_{il1} = x_{il1}$ follows the Clayton copula with parameter $\rho/(1+\rho)$, so,

$$\begin{aligned} &P(X_{ij1} > a_j, X_{ik1} > a_k | X_{il1} = a_l, V_{ij}, V_{ik}, V_{il}) \\ &\quad = C(\mathcal{F}(a_j | X_{il1} = a_l, V_{ij}, V_{il}; \phi_1, \rho), \mathcal{F}(a_k | X_{il1} = a_l, V_{ik}, V_{il}; \phi_1, \rho); \rho^*), \end{aligned} \tag{5.2.3}$$

where $\rho^* = \rho/(1+\rho)$ and

$$\mathcal{F}(a_j | X_{il1} = a_l, V_{ij}, V_{il}; \phi_1, \rho) = \left. \frac{\partial C(\mathcal{F}(a_j|V_{ij};\phi_1), \mathcal{F}(x_l|V_{il};\phi_1);\rho)}{\partial x_l} \right|_{x_l = a_l}.$$

As a measure of dependence of the age of disease onset between two individuals, we consider the cross ratio for (X_{ij1}, X_{ik1}) (Oakes, 1989) which takes the form of

$$\begin{aligned} \theta(a_j, a_k) &= \frac{\lambda_1(a_k | X_{ij1} = a_j; \boldsymbol{V}_i, \varphi)}{\lambda_1(a_k | X_{ij1} > a_j; \boldsymbol{V}_i, \varphi)} \\ &= \frac{P(X_{ij1} = a_j, X_{ik1} = a_k; \boldsymbol{V}_i, \varphi) P(X_{ij1} > a_j, X_{ik1} > a_k; \boldsymbol{V}_i, \varphi)}{P(X_{ij1} = a_j, X_{ik1} > a_k; \boldsymbol{V}_i, \varphi) P(X_{ij1} > a_j, X_{ik1} = a_k; \boldsymbol{V}_i, \varphi)}. \end{aligned} \tag{5.2.4}$$

We note that under the Clayton copula, $\theta(a_j, a_k) = 1 + \rho$. We assume that the (possibly latent) age at disease-free death for an individual is independent from the life history of other family members. This assumption may not be valid in settings where the occurrence of death might be affected by shared environmental factors in a family. Nevertheless we adopt it here, recognizing that there is a within-family dependence in the marginal time of death ($\min(X_{ij2}, X_{ij3})$) accommodated. Note that under the assumption of i) conditionally independent competing risks, $X_{ij1} \perp X_{ij3}|V_{ij}$, and ii) $X_{ij3} \perp \{Z_{ik}(s), 0 < s\}|\boldsymbol{B}_i, \boldsymbol{V}_i$ for $j \neq k$, a cause-specific cross-ratio $\theta_{11}(a_j, a_k)$ for the age of disease onset between two individuals is the same as the cross odds ratio $\theta(a_j, a_k)$ in (5.2.4). This is true since $P(X_{ij1} = a_j, X_{ij3} > a_j, X_{ik1} = a_k, X_{ik3} > a_k; \boldsymbol{B}_i, \boldsymbol{V}_i, \varphi) = P(X_{ij1} = a_j, X_{ik1} = a_k; \boldsymbol{V}_i, \varphi)$

and

$$\theta_{11}(a_j, a_k) = \frac{\lambda_{11}(a_k|X_{ij1} = a_j, X_{ij3} > a_j; \boldsymbol{B}_i, \boldsymbol{V}_i, \varphi)}{\lambda_{11}(a_k|X_{ij1} > a_j, X_{ij3} > a_j; \boldsymbol{B}_i, \boldsymbol{V}_i, \varphi)} = \frac{\lambda_1(a_k|X_{ij1} = a_j; \boldsymbol{V}_i, \varphi)}{\lambda_1(a_k|X_{ij1} > a_j; \boldsymbol{V}_i, \varphi)},$$

where $\theta_{11}(\cdot)$ represents the cause-specific hazard ratio for the transition from 0 to 1 for two members, $\lambda_1(a_k|X_{ij1} = a_j, X_{ij3} > a_j; \boldsymbol{B}_i, \boldsymbol{V}_i, \varphi)$ is the conditional hazard at disease age a_k given the other family member has disease at age a_j, and $\lambda_1(a_k|X_{ij1} > a_j, X_{ij3} > a_j; \boldsymbol{B}_i, \boldsymbol{V}_i, \varphi)$ is the conditional hazard at disease age a_k given the other family member is in healthy state at age a_j under the semi-competing risks setting. Here, we consider the Clayton copula and we let $\theta(a_j, a_k) = \theta_{11}(a_j, a_k) = \theta = 1 + \rho$. Note that since the age of disease onset between family members are correlated, X_{ij2} may not be independent of X_{ik2}.

Scheike *and others* (2010) introduced a cross-odds ratio for the age of disease onset under the competing risks setting given by

$$\pi(a) = \frac{ODDS(X_{ik1} \leq a, X_{ik1} < X_{ik3}|X_{ij1} \leq a, X_{ij1} < X_{ij3}; \boldsymbol{B}_i, \boldsymbol{V}_i)}{ODDS(X_{ik1} \leq a, X_{ik1} < X_{ik3}, B_{ik}, V_{ik})},$$

where

$$P(X_{ik1} \leq a, X_{ik1} < X_{ik3}, B_{ik}, V_{ik}) \tag{5.2.5}$$

is the marginal cumulative incidence function for the age of disease onset. Note that $\pi(a)$ does not have a simple expression even with $\theta(a, a) = 1 + \rho$ under the Clayton copula, because the cumulative incidence functions are obtained by the cause-specific hazards $\lambda_1(\cdot), \lambda_2(\cdot)$, and $\lambda_3(\cdot)$.

5.2.2 Likelihood Construction for Family Studies

We now extend the model to deal with all members of a family. We let $m_i + 1$ denote the total number of members of family i with the subscript 0 used to denote the proband in the family. Let $\boldsymbol{X}_{i1} = (X_{i01}, X_{i11}, \ldots, X_{im_i1})'$ denote the vector of possibly latent onset times within family i, $\boldsymbol{X}_{i3} = (X_{i03}, X_{i13}, \ldots, X_{im_i3})'$, $\boldsymbol{X}_{i2} = (X_{i02}, X_{i12}, \ldots, X_{im_i2})'$, $\boldsymbol{B}_i = (B_{i0}, \ldots, B_{im_i})'$, and $\boldsymbol{V}_i = (V_{i0}, \ldots, V_{im_i})'$. Then (5.2.2) extends to m_i+1 dimensional

survival copula function as

$$P(X_{i01} > a_0, \ldots, X_{im_i1} > a_{m_i} | \boldsymbol{V}_i; \varphi) = \mathcal{C}(\mathcal{F}(a_0|V_{i0}; \phi_1), \ldots, \mathcal{F}(a_{m_i}|V_{im_i}; \phi_1); \rho).$$

We consider studies in which families are sampled by the selection of the proband from a disease registry. We label the proband $j = 0$ and selected family members by $j = 1, \ldots, m_i, i = 1, \ldots, n_F$. We let R_{i0} denote the calendar time of screening and recruitment of the proband to the registry and $C_{i0} = R_{i0} - B_{i0}$ the age of the proband at calendar time R_{i0}. To be in the registry, the proband must be alive with disease at age C_{i0}. Let R_i be the calendar time of the second stage of sampling of the proband from the registry for inclusion in the family study, and A_{i0}, and $\boldsymbol{A}_i = (A_{i0}, A_{i1}, \ldots, A_{im_i})'$ denote the age at calendar time R_i for the proband and all family members, respectively; let $\boldsymbol{A}_i^- = (A_{i1}, \ldots, A_{im_i})'$ denote the elements of $\boldsymbol{A}_i$ excluding the proband. More generally a superscript "-" denotes a vector with the entry for the proband excluded.

Given the life history of the proband, we obtain data from the non-probands. If a non-proband died before R_i, it is often possible to obtain disease history data from medical records or via the proband. Anderson (1961) compared the accuracy of reports about disease histories of family members with diagnosis data from physicians, and found that obtaining information from physicians is necessary to ensure accurate reporting is made for non-probands. We therefore also consider designs in which physicians must interview non-probands at calendar time R_i to carry out medical examinations. In this second design, non-probands would have to be alive at the family recruitment time R_i.

The Lexis diagram plays a central role in describing the incidence, path, and sampling of disease processes in a population using a calendar time $\times$ age co-ordinate system (Keiding, 1990, 2006). Figure 5.2 shows possible scenarios for family data on illness-death processes under the biased sampling scheme. In this figure, the dashed lines represent periods of calendar time and ages at which the healthy state is occupied, and the solid lines represent periods in which the diseased state is occupied. The proband, depicted in red, provides their retrospectively recorded age of disease onset, and like other individuals in the registry may be followed until death or censoring. Non-probands may give a mixed type of data. Some may give retrospectively reported ages of disease onset, some may be disease-free at

the time of examination, and for some we may simply know their date of death if they did not live long enough to be recruited and examined at the calendar time of the family study.

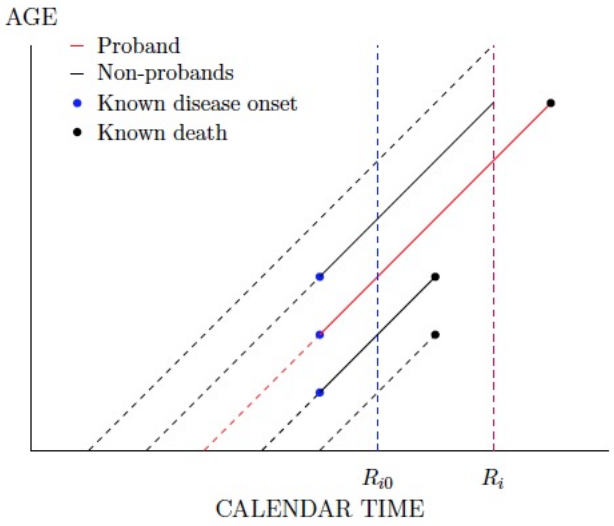


Figure 5.2: A Lexis diagram for family data obtained under a biased sample scheme; R_{i0} denotes the calendar time of recruitment of a proband to a registry and R_i is the date of the family study.

We construct the likelihood function for two particular study designs under the biased sampling schemes, depending on whether we collect all history of non-probands at R_i or only examine non-probands who are alive at R_i. If $\boldsymbol{s}_i = (s_{i0}, s_{i1}, \ldots, s_{im_i})'$ denotes a vector of ages of individuals in family i, we let $\boldsymbol{Z}_i(\boldsymbol{s}_i) = (Z_{i0}(s_{i0}), Z_{i1}(s_{i1}), \ldots, Z_{im_i}(s_{im_i}))'$. In both designs the likelihood contribution of the proband is

$$L_{i0}(\phi) = P(\bar{Z}_{i0}(A_{i0})|Z_{i0}(C_{i0}) = 1, C_{i0}, B_{i0}, V_{i0}; \phi), \tag{5.2.6}$$

where $\bar{\boldsymbol{Z}}_{i0}(A_{i0}) = \{Z_{i0}(u), 0 < u \leq A_{i0}\}$. In the first design we suppose the disease history and covariates for all non-probands are available at calendar time R_i at which the family study is conducted. The likelihood is then given as

$$L_i^I(\psi) \propto L_{i0}(\phi) P(\bar{\boldsymbol{Z}}_i^-(\boldsymbol{A}_i^-)|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{V}_i; \psi), \tag{5.2.7}$$

where $\bar{\boldsymbol{Z}}_i(s_i) = \{Z_{ij}(u), 0 < u \leq s_{ij}, j = 0, \ldots, m_i\}$ and $\bar{Z}_{ij}(s) = \{Z_{ij}(u), 0 < u \leq s\}$.

In the second design we consider a study in which non-probands must be alive to be examined and participate in the family study. This gives

$$L_i^{II}(\psi) \propto L_{i0}(\phi) P(\bar{\boldsymbol{Z}}_i^-(\boldsymbol{A}_i^-) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{Z}_i^-(\boldsymbol{A}_i^-) \in \{0,1\}^{m_i}, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{V}_i; \psi). \tag{5.2.8}$$

In what follows we omit the superscript I and II indicating the design and take it as understood that L_i represents either L_i^I or L_i^{II} in a particular setting. The score vector is $S(\psi) = \sum_{i=1}^{n_F} S_i(\psi) = \sum_{i=1}^{n_F} \partial \log L_i / \partial \psi$ and the information matrix is

$$I(\psi) = -\sum_{i=1}^{n_F} I_i(\psi) = -\sum_{i=1}^{n_F} \partial^2 \log L_i(\psi) / \partial \psi \partial \psi',$$

respectively. We obtain the maximum likelihood estimator $\hat{\psi}$ by solving $S(\psi) = 0$ and asymptotically $\sqrt{n_F}(\hat{\psi} - \psi) \sim N(0, \mathcal{I}^{-1}(\psi))$ where $\mathcal{I}(\psi) = E[I_i(\psi)]$.

When m_i is large the computational burden of evaluating the joint probability of the life histories of family members may be considerable, so we consider use of "pairwise" conditional composite likelihood (Varin *and others*, 2011) in which pairs are comprised of two non-probands and the models condition on the proband data for this respective family. In particular, in the second design in (5.2.8) where non-probands are only selected if they are alive, the composite likelihood is exploited to address the difficulty of calculating the condition $P(\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{Z}_i^-(\boldsymbol{A}_i^-) \in \{0,1\}^{m_i}, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{V}_i)$. Then, the contribution to the conditional composite likelihood of family i for design k is

$$CL_i^k(\psi) \propto L_{i0}(\phi) \prod_{1 \leq j < l \leq m_i} \left\{ L_{ijl}^k(\psi) \right\}^{\frac{1}{m_i - 1}}, \qquad k = I, II, \tag{5.2.9}$$

where L_{i0} is given by (5.2.6); the weight $1/(m_i - 1)$ ensures the contribution to the marginal function for non-probands is appropriate. In design I,

$$L_{ijl}^I(\psi) = P(\bar{\boldsymbol{Z}}_{ijl}^-(\boldsymbol{A}_{ijl}^-) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$

and in design II,

$$L_{ijl}^{II}(\psi) = P(\bar{\boldsymbol{Z}}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-})|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{Z}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}) \in \{0,1\}^2, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$
$$= \frac{P(\bar{\boldsymbol{Z}}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-})|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)}{P(\boldsymbol{Z}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}) \in \{0,1\}^2|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)}$$

with $\boldsymbol{A}_{ijl} = (A_{i0}, A_{ij}, A_{il})'$, $\boldsymbol{B}_{ijl} = (B_{i0}, B_{ij}, B_{il})'$, $\boldsymbol{V}_{ijl} = (V_{i0}', V_{ij}', V_{il}')'$, $\bar{\boldsymbol{Z}}_{ijl}(s_{ijl}) = \{Z_{ih}(u), 0 < u \leq s_{ih}, h = 0, j, l; \boldsymbol{B}_{ijl}\}$. We calculate $L_{ijl}^{k}(\psi)$ using the conditional Clayton copula function based on (5.2.3). The detail of computation of composite likelihood is given in Appendix 5.A. Again we suppress the superscript I or II when discussing a generic setting we write $CL_i(\psi)$. The score vector for the composite likelihood is then $U(\psi) = \sum_{i=1}^{n_F} U_i(\psi) = \sum_{i=1}^{n_F} \partial \log CL_i(\psi)/\partial\psi =$ and the maximum composite likelihood estimator $\tilde{\psi}$ is obtained by solving $U(\psi) = 0$. The estimated variance of $\tilde{\psi}$ is given as $n_F^{-1}A^{-1}(\tilde{\psi})B(\tilde{\psi})A^{-1}(\tilde{\psi})$ where $A(\psi) = -n_F^{-1}\sum_{i=1}^{n_F} \partial U_i(\psi)/\partial\psi'$, and $B(\psi) = n_F^{-1}\sum_{i=1}^{n_F} U_i(\psi)U_i'(\psi)$.

5.3 Augmented Composite likelihood

In the context of family study, the low incidence of disease onset among non-probands and bias sampling scheme pose difficulty for analysis due to the lack of data. To overcome this difficulty, auxiliary data can provide additional source of data and strengthen the analysis. The combination of data from different sources in family study has been suggested (Pfeiffer *and others*, 2008; Zheng *and others*, 2010; Balliu *and others*, 2012) in which the case-control studies or the twin-based studies are incorporated with the family-based studies. In our motivating example, University of Toronto Psoriatic Arthritis Registry (UTPAR) provides data with right-truncated disease onset time and the left-truncated and right-censored time to death (Wong *and others*, 1997). The UTPAR also conducts tracing studies, which aim to yield further data on survival times for PsA patients. Another source of auxiliary data is a national cross-sectional survey conducted by the National Psoriasis Foundation in the United States; it yields current disease status data (Gelfand *and others*, 2005). Although this study provides only marginal information, the efficiency can be enhanced

by augmenting the likelihood. Since we have no data available on the time to disease-free death, we use national mortality statistics to estimate the disease-free mortality rate; the data are population-level data and so we treat $\lambda_3(\cdot,\cdot)$ as known and define them to be the population mortality rates. We thus consider i) registry data with follow-up ii) a cross-sectional survey yielding current status data on disease state, and iii) national statistics for the mortality rate.

Let $\mathcal{A}_1$ the set of individuals in the registry, $\mathcal{A}_2$ the set of individuals from the cross-sectional survey. We multiply $CL^k(\psi)$ in (5.2.9) by the corresponding marginal likelihood $L_{\mathcal{A}_1}, L_{\mathcal{A}_2}$ based on auxiliary data i), ii), respectively. For the individuals in the registry except the probands in i), we let X_{r1} denote the age at onset, C_r the age at recruitment, X_{r2} the age at death following disease (if available), $A_r^* = \min(C_r^*, X_{r2})$ with C_r^* the last assessment time, B_r the calendar time of birth, and V_r a vector of covariates for an individual r. Then, $L_{\mathcal{A}_1}$ is given as

$$L_{\mathcal{A}_1} \propto \prod_r^{n_R} P(\bar{Z}_r(A_r^*)|Z_r(C_r) = 1, C_r, B_r, V_r; \phi)$$

where n_R is a size of registry data.

In the case of ii), if C_r denotes the age at contact for the survey, then $L_{\mathcal{A}_2}$ is written as

$$L_{\mathcal{A}_2} \propto \prod_r^{n_S} \prod_{h\in\{0,1\}} P(Z_r(C_r) = h|Z_r(C_r) \in \{0,1\}, B_r, V_r; \phi)^{I(Z_r(C_r)=h)}$$

where n_S is the sample size of the survey. In the case of iii), we obtain $\lambda_3(t,a)$ using the published population data which is calculated by the number of deaths in age-, calendar time- divided by the exposure-to-risk in age-, calendar time- (Robert, 2017). Figures 5.3 shows the age-specific population mortality rates across calendar periods between 1921 to 2011. A decreasing trend in the age-specific mortality rates over the last 90 years is apparent, therefore, if the registry includes individuals born over a wide range of calendar time, age- and time-specific mortality rates should be considered.

To examine the asymptotic distribution of the estimator $\tilde{\psi}$, we construct the augmented

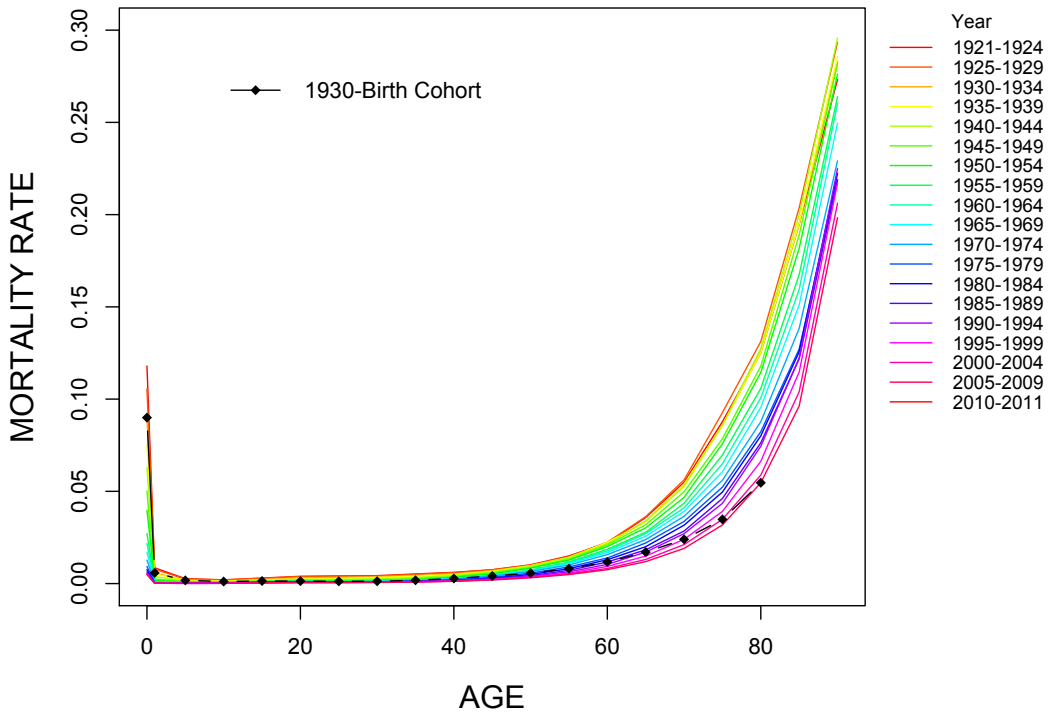


Figure 5.3: Age-specific population mortality rates by calendar period in Canada from 1921 to 2011

composite likelihood

$$ACL(\psi) \propto \prod_{i=1}^{n_F} CL_i(\psi) \prod_{r=1}^{n_R} L_{\mathcal{A}_1,r}(\phi) \prod_{r=1}^{n_S} L_{\mathcal{A}_2,r}(\phi), \qquad k = I, II, \tag{5.2.10}$$

and we may write

$$U_{\mathcal{F},i}(\psi) = \frac{\partial \log CL_i(\psi)}{\partial \psi},$$

$$U_{\mathcal{A}_1,r}(\phi) = \frac{\partial \log L_{\mathcal{A}_1,r}(\phi)}{\partial \psi},$$

and

$$U_{\mathcal{A}_2,r}(\phi) = \frac{\partial \log L_{\mathcal{A}_2,r}(\phi)}{\partial \psi}.$$

The score vector for the augmented composite likelihood is

$$\bar{U}(\psi) = \sum_{i=1}^{n_F} U_{\mathcal{F},i}(\psi) + \sum_{r=1}^{n_R} U_{\mathcal{A}_1,r}(\phi) + \sum_{r=1}^{n_S} U_{\mathcal{A}_2,r}(\phi)$$

and the maximum augmented pairwise likelihood estimator $\tilde{\psi}$ is obtained by solving $\bar{U}(\psi) = 0$. The estimated variance of $\tilde{\psi}$ is given as $n^{-1}A^{-1}(\tilde{\psi})B(\tilde{\psi})A^{-1}(\tilde{\psi})'$ where

$$A(\psi) = -\frac{1}{n}\left(\sum_{i=1}^{n_F} \frac{\partial^2 \log CL_i(\psi)}{\partial\psi\partial\psi'} + \sum_{r=1}^{n_R} \frac{\partial^2 \log L_{\mathcal{A}_1,r}(\phi)}{\partial\psi\partial\psi'} + \sum_{r=1}^{n_S} \frac{\partial^2 \log L_{\mathcal{A}_2,r}(\phi)}{\partial\psi\partial\psi'}\right),$$

and

$$B(\psi) = \frac{1}{n}\left(\sum_{i=1}^{n_F} U_{\mathcal{F},i}(\psi)U'_{\mathcal{F},i}(\psi) + \sum_{r=1}^{n_R} U_{\mathcal{A}_1,r}(\phi)U'_{\mathcal{A}_1,r}(\phi) + \sum_{r=1}^{n_S} U_{\mathcal{A}_2,r}(\phi)U'_{\mathcal{A}_2,r}(\phi)\right)$$

with $n = n_F + n_R + n_S$.

5.4 Simulation Studies

Here we assess the performance of the methods introduced in Section 5.2 and 5.3 through simulation studies. To mimic more closely the PsA study, we consider the age and calendar time-specific mortality rates based on the population mortality rates $\lambda_3(t, a)$ and assume $\lambda_2(t, a) = \nu\lambda_3(t, a)$. We set the rate of occurrence of disease $\lambda_1 = 0.01$ as a constant value. We consider the Clayton copula with Kendall's $\tau = 0.2$ and 0.4. We generate the time to disease-free death from the age and time-specific population mortality rates with $\nu = 1.1$. We generate the family size with 4 or 6 members having two parents and 2 or 4 children in family where $P(m_i + 1 = 4) = 2/3$ and $P(m_i + 1 = 6) = 1/3$. Then we randomly choose an individual from the family members and generate the date of birth from the uniform distribution $(1920, 1950)$ if the individual is a parent or $(1950, 1980)$ otherwise. Then we

generate an individual path from the marginal distribution. We generate the individual sampling date from the uniform distribution (1980, 2010) and select those who are alive and diseased at the sampling date and repeat for the registry with the family sample size $n_R + n_F$. Among $n_R + n_F$ individuals who are alive at the family sampling date on July 1st of 2010, we randomly select probands and generate the data for non-probands given the proband data with the family size n_F. If the proband is a parent, the birth dates of spouse or children are obtained by adding the uniform distribution $(0, 10)$ or $(20, 30)$ to the birth date of proband, respectively, and conduct similarly when the proband is a child. In design I, we include all non-probands data in analysis, whereas we only include alive non-probands data in design II. In this simulation, we consider both types of auxiliary data: the registry data with follow-up and the current status survey data. The registry follow-up data including probands are assumed to be collected until July 1st of 2010 with the record of death post disease. For the current status survey data, we generate the date of birth from the uniform distribution $(1930, 1980)$ and set the sampling date as July 1st of 2000. We set the family sample size $n_F = 1000$, the size of registry $n_R = 2000$, and the survey size $n_S = 1000$. Here, the augmented pairwise estimations were carried out and the results are reported in Table 5.1 for design I, and Table 5.2 for design II, respectively.

For all methods in two designs, the biases are negligible, the empirical standard errors (ESEs) are in a good agreement with the average standard errors (ASEs), and the empirical coverage probability (ECP) of nominal 95% confidence intervals are all within an acceptable range. The estimators under the full likelihood have smaller ASEs compared to the pairwise likelihood with the registry data; however, the current status auxiliary data improve efficiency so that the estimators obtained by the pairwise likelihood are as efficient as those by the full likelihood. Since the current status auxiliary data have no time to death data, the efficiency of ν is not improved. Comparing design I and II, the estimators have better efficiency under design I. Also, the estimators λ_1 and τ under design II are as nearly as efficient as those under design I with current status data.

Table 5.1: Frequency properties of estimators based on the augmented pairwise likelihood for family data given $\lambda_3(\cdot,\cdot)$ under biased sampling scheme for the proband and disease history of non-probands available (design I) with two auxiliary data: the registry follow-up data and the current status survey data; Clayton copula with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_R = 2000$, $n_S = 1000$, and $nsim = 1000$

		Registry Data				Registry + Current Status Data			
τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
				Full Likelihood					
0.2	$\log(\lambda_1)$	-0.003	0.057	0.058	0.951	-0.001	0.040	0.040	0.948
	$\log(\nu)$	-0.001	0.038	0.037	0.947	-0.001	0.038	0.037	0.950
	τ	0.001	0.028	0.028	0.947	-0.000	0.022	0.022	0.956
0.4	$\log(\lambda_1)$	-0.003	0.079	0.080	0.960	-0.000	0.044	0.045	0.954
	$\log(\nu)$	-0.002	0.037	0.036	0.952	-0.002	0.037	0.036	0.948
	τ	0.002	0.033	0.033	0.950	0.001	0.021	0.021	0.953
				Pairwise Likelihood					
0.2	$\log(\lambda_1)$	-0.003	0.059	0.061	0.953	-0.001	0.040	0.041	0.959
	$\log(\nu)$	-0.001	0.038	0.037	0.943	-0.001	0.038	0.037	0.946
	τ	0.001	0.030	0.030	0.950	-0.000	0.023	0.023	0.961
0.4	$\log(\lambda_1)$	-0.004	0.081	0.082	0.963	-0.001	0.044	0.045	0.954
	$\log(\nu)$	-0.002	0.037	0.036	0.951	-0.002	0.037	0.036	0.951
	τ	0.002	0.035	0.034	0.957	0.001	0.022	0.021	0.962

Table 5.2: Frequency properties of estimators based on the augmented pairwise likelihood for family data given $\lambda_3(\cdot,\cdot)$ under biased sampling scheme for the proband and alive non-probands data available (design II) with two auxiliary data: the registry follow-up data and the current status survey data; Clayton copula with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_R = 2000$, $n_S = 1000$, and $nsim = 1000$

		Registry Data				Registry + Current Status Data			
τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
0.2	$\log(\lambda_1)$	-0.006	0.065	0.066	0.949	-0.001	0.042	0.042	0.948
	$\log(\nu)$	-0.001	0.048	0.046	0.957	-0.001	0.048	0.046	0.940
	τ	0.003	0.033	0.033	0.951	0.000	0.024	0.025	0.952
0.4	$\log(\lambda_1)$	-0.002	0.085	0.086	0.962	0.000	0.045	0.046	0.957
	$\log(\nu)$	-0.002	0.048	0.046	0.937	-0.002	0.046	0.047	0.938
	τ	0.002	0.036	0.037	0.950	0.002	0.023	0.022	0.958

5.5 Assessment of Genetic Risk Factors

If familial aggregation is identified by the proposed model in Section 5.2 and 5.3, interest may lie in the effect of genetic factors on disease onset to explain familial aggregation. However, if some individuals in the study are not genotyped, incomplete genetic data must be dealt with. For example, in design I, we may obtain the disease history for non-probands who died but cannot sample their DNA. Also, the national current status survey data do not provide the genetic information. Chatterjee *and others* (2006) proposed an analysis for a kin-cohort case-control and case-only family data with genotype and phenotype. Gong *and others* (2010) categorized two family designs: the population and the clinic designs and present the simulation studies to examine the performance of phenotype/genotype-based methods. Zhang *and others* (2010) suggested statistical methods in estimating age-dependent penetrance under a case-family design.

In this section, we accommodate genetic data in our model but deal with missing genetic information. We let G_{ij} denote the genotype (gene carrier indicator), which is

tentatively related to disease with $P(G_{ij}=0)=q^2, P(G_{ij}=1)=p^2+2pq$ with the allele frequency p and $q=1-p$ for individual j in family i and $\boldsymbol{G}_i=(G_{i0},\ldots,G_{im_i})'$. We denote $W_{ij}=(V_{ij}',G_{ij})'$ a vector of covariates and genotypes and $\boldsymbol{W}_i=(\boldsymbol{V}_i',\boldsymbol{G}_i')'$. The transition intensities, then, are written as $\lambda_l(t,a|b_{ij},w_{ij})$ for $l=1,2,3$ where v_{ij} is replaced with w_{ij}. The joint probability of disease onset also needs to replace $\boldsymbol{V}_i$ with $\boldsymbol{W}_i$ but the cross-ratio or cause-specific hazard ratio under the Clayton copula remains the same as θ. We make the following additional assumptions: i) The process is in Hardy-Weinberg equilibrium and the Mendelian law holds, ii) $G_{ij}\perp V_{ij}$, iii) $\bar{Z}_{ij}(s)|G_{ij}\perp G_{ik}\ \forall s$ for $j\neq k$, iv) $\lambda_1(t,a|b_{ij},w_{ij})=\lambda_1(a)\exp(g_{ij}\alpha+v_{ij}'\beta_1)$, and v) $\lambda_2(t,a|b_{ij},w_{ij})=\lambda_2(t,a|b_{ij},v_{ij})$ and $\lambda_3(t,a|b_{ij},w_{ij})=\lambda_3(t,a|b_{ij})$.

5.5.1 Composite Likelihood with Incomplete Genetic Data

Here, we focus on the augmented pairwise conditional likelihood in Section 5.3. First, we consider design II with two sources of auxiliary data: i) the family study data and the registry data and ii) the family study data, the registry data, and current status data from the survey. In the former case, all individuals are genotyped in the family study and the registry since they are all examined, so we can assume that the genotypes are given and the pairwise composite likelihood does not change the form of likelihood which has the genotype variable as a covariate. However, the genotype data are not available in the survey, so in the latter setting, we need to model G_{ij}. The contribution of the proband to the likelihood is then

$$\begin{aligned} L_{i0}(\phi) &= P(\bar{Z}_{i0}(A_{i0}),G_{i0}|Z_{i0}(C_{i0})=1,C_{i0},B_{i0},V_{i0};\phi) \\ &= \frac{P(\bar{Z}_{i0}(A_{i0})|G_{i0},C_{i0},B_{i0},V_{i0};\phi)P(G_{i0})}{\sum_{g\in(0,1)}P(Z_{i0}(C_{i0})=1|C_{i0},B_{i0},G_{i0}=g,V_{i0};\phi)P(G_{i0}=g)}, \end{aligned} \tag{5.2.11}$$

where we select the proband based only on phenotype (disease status) at R_{i0}. Then the contribution from the non-probands $L_i^{II}(\psi)$ is

$$L_{ijl}^{II}(\psi) = P(\bar{\boldsymbol{Z}}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}), \boldsymbol{G}_{ijl}^{-} | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, G_{i0}, \boldsymbol{Z}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}) \in \{0,1\}^2, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$
$$= \frac{P(\bar{\boldsymbol{Z}}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{G}_{ijl}, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi) P(\boldsymbol{G}_{ijl}^{-} | G_{i0})}{\sum_{g \in \{0,1\}^2} P(\boldsymbol{Z}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}) \in \{0,1\}^2 | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, G_{i0}, \boldsymbol{G}_{ijl}^{-} = g, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi) P(\boldsymbol{G}_{ijl}^{-} = g | G_{i0})}$$

where $\boldsymbol{G}_{ijl} = (G_{i0}, G_{ij}, G_{il})'$ and $P(\boldsymbol{G}_{ijl}^{-} | G_{i0})$ can be calculated using the allele frequency f and family structure; see Appendix 5.B. For the auxiliary data, we let G_r denote the genotype of individual r in $\mathcal{A}_1$ or $\mathcal{A}_2$ in Section 5.3. The likelihood terms based on the auxiliary data $L_{\mathcal{A}_1}$ and $L_{\mathcal{A}_2}$ are then given as

$$L_{\mathcal{A}_1} \propto \prod_{r}^{n_R} P(\bar{Z}_r(A_r^*), G_r | Z_r(C_r) = 1, C_r, B_r, V_r),$$

and

$$L_{\mathcal{A}_2} \propto \prod_{r}^{n_S} \prod_{h \in \{0,1\}} \Big\{ \sum_{g} P(Z_r(C_r) = h | Z_r(C_r) \in \{0,1\}, B_r, G_r = g, V_r) \times P(G_r = g | Z_r(C_r) \in \{0,1\}, B_r, V_r) \Big\}^{I(Z_r(C_r)=h)}$$

where

$$P(G_r = g | Z_r(C_r), B_r, V_r) = \frac{P(Z_r(C_r) | G_r = g, B_r, V_r) P(G_r = g)}{\sum_{g \in \{0,1\}} P(Z_r(C_r) | G_r = g, B_r, V_r) P(G_r = g)}.$$

Secondly, we only observe the genotype of non-probands who are alive in design I, and non-probands who did not survive to R_i are not genotyped. In this case

$$L_{ijl}^{I}(\psi) = P(\bar{\boldsymbol{Z}}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}), \boldsymbol{G}_{ijl}^{o-} | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, G_{i0}, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$
$$= \sum_{\boldsymbol{G}_{ijl}^{m}} P(\bar{\boldsymbol{Z}}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{G}_{ijl}, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi) P(\boldsymbol{G}_{ijl}^{-} | G_{i0})$$

where $\boldsymbol{G}_{ijl}^{o}$ is a vector of observed genotype in family i for the pair of family member j and l with the proband genotype on the first component, $\boldsymbol{G}_{ijl}^{m}$ is a vector of missing genotypes for the family member j and l in family i, and $\boldsymbol{G}_{ijl} = (\boldsymbol{G}_{ijl}^{o'}, \boldsymbol{G}_{ijl}^{m'})'$.

5.5.2 Simulation Studies

We conducted further simulation studies to assess performance of the proposed model with genetic risk factors. We considered a binary indicator G_{ij} with the allele frequency $p = 0.06$ with a hazard ratio $= \exp(\alpha) = 1.5$; we do not consider additional covariates for simplicity and otherwise adopt the same simulation settings as in Section 5.4.

We first generate the genotype for family members based on the family structure under the Mendelian law and given the genotype we generate family members' lifetime paths based on the proposed model. The selection criteria remains the same as in Section 5.4. The empirical properties of the estimators for the parameters based on design I and II are reported in Table 5.3 and 5.4, respectively.

Here we can observe the same findings pointed out in Section 5.5. The current status survey data do not affect the efficiency α and p because the genetic marker is not available in the survey, however, they increase the efficiency of λ_1 and Kendall's τ in design I. This highlights the value of the current status data when disease onset times are right-truncated even for the dependence parameter. In design II, the current status data improve efficiency of each estimator except the one for ν. It is therefore advantageous for score tests, in particular, when interest lies in testing genetic effects on disease onset as it may increase the power of such tests.

5.6 Application to the Psoriatic Arthritis Family Study

Psoriasis is an inflammatory skin disease occurring about 2-3% of the general population and PsA(Psoriatic Arthrists) is an inflammatory arthritis disease affecting about 30% of patients with psoriasis (Gladman, 1991; Langley *and others*, 2005; Eder *and others*, 2012). Patients with PsA are at higher risk for death compared to the general population of

Table 5.3: Frequency properties of estimators based on the augmented pairwise likelihood for family data with genotype information given $\lambda_3(\cdot,\cdot)$ under biased sampling scheme for the proband and disease history of non-probands available (design I) with two auxiliary data: the registry follow-up data and the current status survey data; Clayton copula with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_R = 2000$, $n_S = 1000$, and $nsim = 1000$

		Registry Data				Registry + Current Status Data			
τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
0.2	$\log(\lambda_1)$	0.002	0.057	0.059	0.952	-0.000	0.041	0.040	0.954
	α	-0.003	0.064	0.064	0.948	-0.002	0.064	0.064	0.947
	$\log(\nu)$	-0.001	0.037	0.037	0.956	-0.001	0.037	0.037	0.957
	$\log(p)$	0.002	0.055	0.056	0.947	0.002	0.055	0.055	0.947
	τ	-0.000	0.028	0.029	0.963	0.001	0.023	0.023	0.953
0.4	$\log(\lambda_1)$	0.002	0.076	0.078	0.950	0.000	0.045	0.044	0.952
	α	-0.002	0.057	0.058	0.948	-0.002	0.058	0.058	0.945
	$\log(\nu)$	-0.001	0.035	0.036	0.952	-0.001	0.035	0.036	0.946
	$\log(p)$	0.002	0.053	0.054	0.946	-0.003	0.053	0.053	0.947
	τ	-0.001	0.032	0.033	0.953	-0.000	0.022	0.022	0.949

Table 5.4: Frequency properties of estimators based on the augmented pairwise likelihood for family data with genotype information given $\lambda_3(\cdot,\cdot)$ under biased sampling scheme for the proband and alive non-probands data available (design II) with two auxiliary data: the registry follow-up data and the current status survey data; Clayton copula with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_R = 2000$, $n_S = 1000$, and $nsim = 1000$

		Registry Data				Registry + Current Status Data			
τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
0.2	$\log(\lambda_1)$	0.001	0.061	0.064	0.951	-0.000	0.042	0.042	0.941
	α	-0.003	0.070	0.071	0.953	-0.003	0.065	0.065	0.949
	$\log(\nu)$	-0.001	0.047	0.046	0.951	-0.001	0.046	0.046	0.949
	$\log(p)$	-	-	-	-	0.002	0.055	0.055	0.949
	τ	0.000	0.030	0.032	0.956	0.001	0.024	0.024	0.952
0.4	$\log(\lambda_1)$	0.004	0.081	0.082	0.955	0.001	0.046	0.045	0.951
	α	-0.001	0.063	0.062	0.949	-0.002	0.059	0.058	0.948
	$\log(\nu)$	-0.002	0.047	0.046	0.948	-0.002	0.046	0.046	0.950
	$\log(p)$	-	-	-	-	0.002	0.053	0.053	0.942
	τ	-0.001	0.035	0.035	0.949	0.001	0.023	0.023	0.941

Ontario with a standardised mortality ratio of 1.36 (Gladman, 2008). Many studies showed that psoriasis is a heritable disease; Pedersen *and others* (2008) reported an increased concordance measure in monozygotic relative to dizygotic twins and Chandran *and others* (2009) confirmed a high familial recurrence risk of PsA based on family studies as shown in Moll and Wright (1973). To obtain a better sense of heredity, Gladman and Farewell (1995); Pedersen *and others* (2008); Chandran and Raychaudhuri (2010); Eder *and others* (2012) identified genes related to psoriasis and PsA and explored environmental factors which increase the risk of PsA. We consider the Human Leucocyte Antigens (HLA)- B27, and HLA-C06 by the findings of the genetic aetiology of psoriasis and psoriatic arthritis in the literature.

We consider data from the Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto which recruited the Psoriatic Arthritis Toronto Cohort and among 1436 individuals from the registry, 150 were selected for family studies as probands. In this family studies, family members were recruited to conduct a thorough examination including genotype information, therefore, this study design belongs to the biased sampling scheme design II. To simplify the analysis, we generate a number of 168 pseudo-families from the original 150 families where two-generation families are considered with the non-missing date of birth and genotype information and we use this pseudo-family data. In the pseudo-family data, the family sizes range from 2 to 7 individuals; 56 families have 2 family members (1 proband and 1 non-proband), and 112 families have at least three members. 193 individuals were diagnosed with PsA among a total of 532 individuals. 145 families have one member with PsA (i.e. proband), 21 families having two members with PsA, and 2 families with three PsA patients in their family.

As a source of auxiliary data, we use the survey of US population in which Gelfand *and others* (2005) reported the prevalence of psoriatic arthritis in 2001. In this survey, subjects with 18 years of age or older were randomly selected and provided the status of psoriasis and psoriatic arthritis; 328 have psoriatic arthritis among 15,307 respondents.

We begin with the model not using the genotype information. We fit a marginal model for the age at PsA onset with piecewise constant hazards with a cut-point 40 to distinguish early and late onset of PsA and assume $\lambda_2(t,a) = \lambda_3(t,a)\nu$ and $\lambda_3(t,a)$ is given as the age-, calendar time- population mortality (Robert, 2017). In the registry data, individuals

Table 5.5: Estimates of parameters based on the augmented pairwise likelihood; auxiliary data include the University of Toronto Psoriatic Arthritis Registry and the survey from Gelfand et al. (2005) without/with genotype variable under the Exponential model and piecewise constant marginal model for age at PsA onset with a cut point 40

MARKER	α_{marker}	ν	τ	p_{marker}
-	-	1.152 (0.016)	0.362 (0.083)	-
B27	0.605 (0.239)	1.155 (0.080)	0.345 (0.085)	0.054 (0.012)
C06	0.117 (0.086)	1.155 (0.060)	0.362 (0.089)	0.115 (0.011)

with missing genotype are dealt with similarly in the survey data. Table 5.5 summarizes the estimates of fitted model without genetic variable in the first column followed by two univariate models with genotype HLA-B27, and HLA-C06 variables including the allele frequency p for each genetic markers.

First, based on the model without genetic markers, we find that $\hat{\nu} = 1.152$ indicating that the ratio of the hazard of death post PsA to PsA-free death is 1.152, which is lower than the reported value in Gladman (2008). As expected, PsA is not lethal while it increases the risk of death. The estimate of dependence parameter is $\hat{\tau} = 0.362$ (95% CI: 0.199, 0.525; p$<$ 0.001) which indicates significant association between family members. The cross ratio or the cause-specific cross ratio is $\hat{\theta} = 2.134$ (95% CI: 1.354, 2.914; p $<$ 0.001), corresponding to 2.134 times higher risk of PsA with a family history of PsA. We find that HLA-B27 has a significant effect on PsA onset ($RR = 1.831$; 95% CI: 0.137, 1.073; p=0.011) but HLA-C06 positive is not associated with an increased risk of PsA ($RR = 1.124$; 95% CI: -0.052, 0.286; p=0.174). This findings are reported in literature that HLA-C06 increases the risk of psorisis but PsA (Chandran, 2013). The allele frequency of HLA-B27 is 0.054, which is compatible with the value of 0.061 from the national USA prevalence of HLA-B27 (Reveille *and others*, 2012). HLA-C06 has the allele frequency 0.115 which is more prevalent that HLA-B27. After adjusted significant genetic marker HLA-B27, Kendall's $\hat{\tau}$ decreases to 0.345 (95% CI: 0.178, 0.512; p$<$ 0.001) since HLA-B27 partially explains the residual familial aggregation.

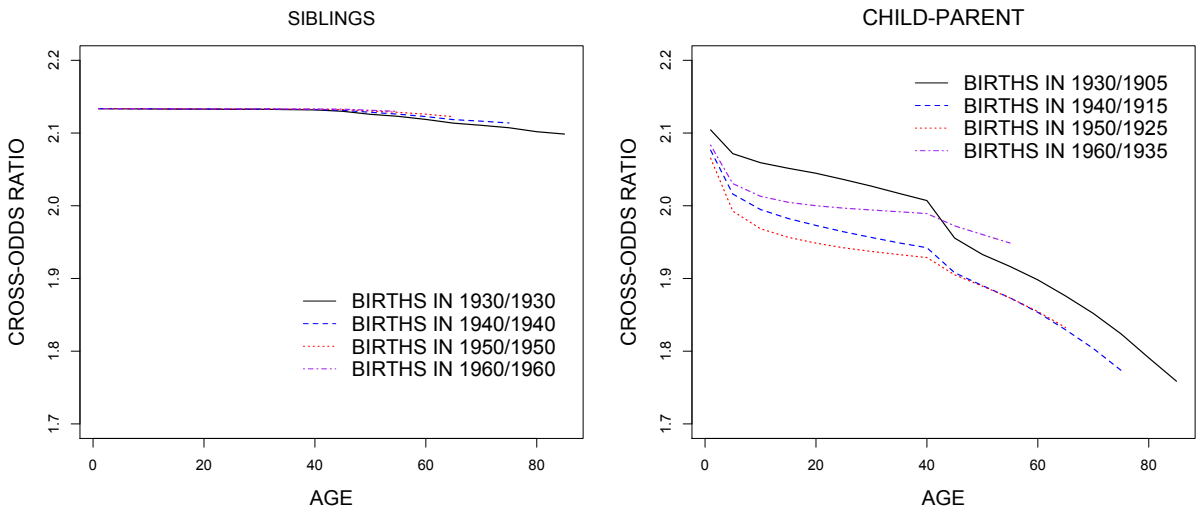


Figure 5.4: The cross-odds ratio for two siblings born in the same year 1930, 1940, 1950, 1960 (the right panel) and a child born in 1930, 1940, 1950, or 1960 given a parent born in 1905, 1915, 1925, or 1935 (the left panel) based on the fitted model with no effect of a genetic marker

Figure 5.4 shows the cross-odds ratio for a sibling given other sibling born in the same year 1930, 1940, 1950, 1960 (the left panel) and a child born in 1930, 1940, 1950, 1960 given a parent born in 1905, 1915, 1925, 1935, respectively. For the sibling pairs, two siblings are governed by the same mortality rates belonging the same birth cohort. The cross-odds ratio before 40 almost plateaus but showed a decreasing trend as they age because the mortality rate increases. There is a drastic decrease in the cross-odds ratio as age increases for the child-parent pairs compared to the sibling pairs. This difference arises due to the higher mortality rates for parents; see Figure 5.3. Similar patterns of the cross-odds ratio for different birth cohorts are observed, but the variation exists.

Figure 5.5 shows the marginal probability of death (state 2 and 3) and the cumulative incidence function for the age of PsA defined in (5.2.5) for different birth cohorts at 1930, 1940, 1950, and 1960. We find that PsA itself is a rare disease with the low cumulative incidence function.

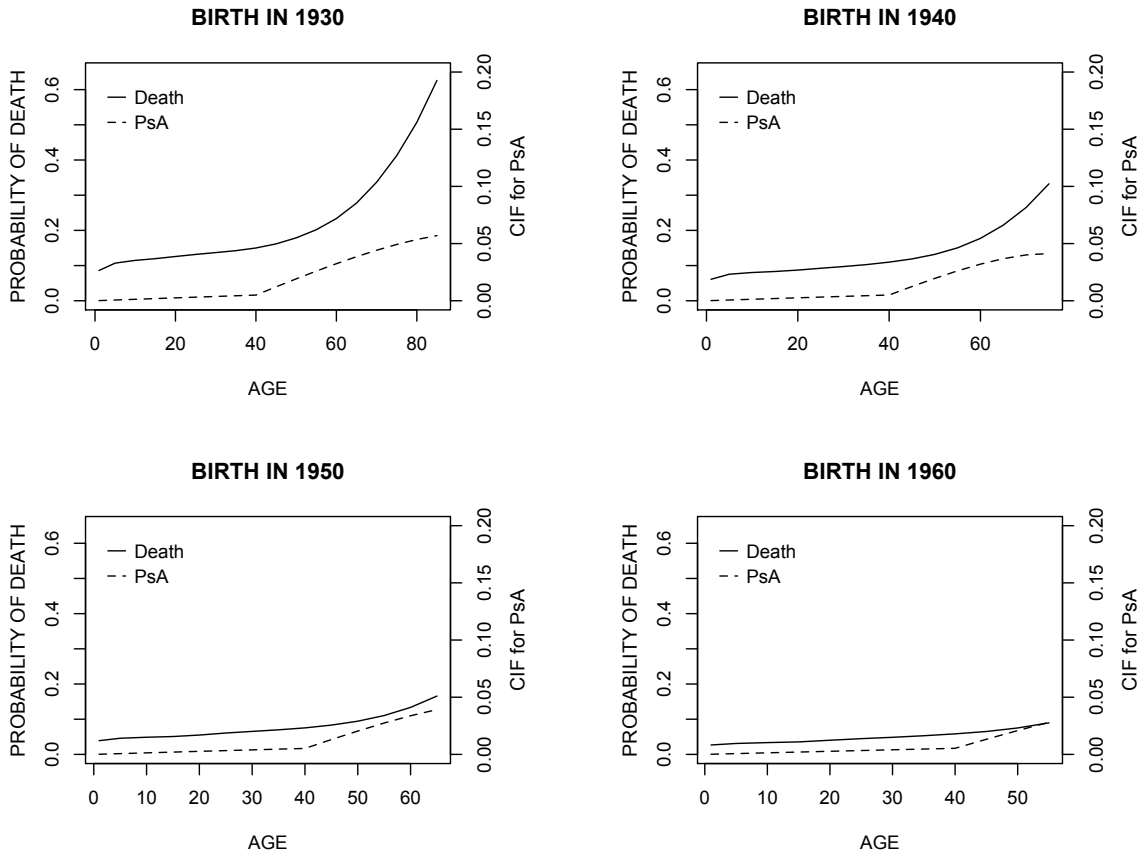


Figure 5.5: The marginal probability of death and the cumulative incidence of PsA by the year of birth of 1930, 1940, 1950, or 1960 based on the fitted model with no effect of a genetic marker

5.7 Discussion

In this Chapter we have proposed an illness-death model for family studies incorporating within-family dependence in the age at disease onset via a copula model The illness-death model offers a natural framework to consider survival bias and the Clayton copula models retain simple interpretations of cross ratio/cause-specific cross ratio and marginal interpretations of estimates of covariance. We explore two study designs for family studies with biased sampling schemes and developed statistical methods for analysis. Pairwise compos-

ite likelihood is utilized to ease the computational burden. We exploit auxiliary data to address identifiability and estimability issues. Age- and calendar time-specific population mortality rates adequately address the trend of mortality rates in family studies where more than two generations are considered. We extend our model to study the effect of genetic markers on risk of disease in which the availability of genotype data depends on the study design. To be more flexible, if the motivating example concerns lethal diseases such as cancer, we allow $\nu(a)$ to be an age-dependent function.

We restrict our attention to the case-only probands family studies. If the case-control probands are available, it would be useful to compare the robustness to misspecification of model assumption (Chatterjee *and others*, 2006) and compare the efficiency with the case-only probands family studies. It is natural to extend our model to allow for different dependence structures in families using a more flexible Gaussian copula (Zhong and Cook, 2016; Lakhal-Chaieb *and others*, 2018). It may also be useful to adopt age-, time-, and sex-specific population mortality rates. As a future work we can examine the effect of ignoring the survival biases arising from the case II design in standard methods of analysis (Zhong and Cook, 2016). The sensitivity of inferences in both standard methods and the proposed methods to within-family dependence in mortality can also be investigated using large sample theory and simulation. As we have shown in the simulation studies in Section 5.3 and 5.5.2, the use of auxiliary data improves efficiency in estimating the marginal parameters related to disease onset and the dependence parameter, so examination of power improvement with auxiliary data would be of interest.

In our motivating example, we focus on the occurrence of psoriatic arthritis. However, PsA occurs in 10-20% of patients with psoriasis and the genetic marker HLA-C06 mostly contributes to develop psoriasis (Queiro *and others*, 2015). To distinguish the genetic risk factors for psoriasis with those for PsA, we may introduce the state of psoriasis in our analysis. Another extension would be to use multiple allele in our analysis. This leads to computational burden due to the summation of all possible combination of genetic markers for missing genotypes. We may use other sources of population studies to calculate the allele frequency and we may exploit this value to assume that the allele frequency p is known in our proposed model. This will reduce the number of parameters to estimate.

Appendix 5.A. An Illustration of Composite Likelihood Construction

Here we give an illustrative of how to construct the composite conditional likelihood, described in Section 5.3. We omit the subscript i labelling family for simplicity and consider no covariates. We consider a particular family consisting of two parents and two children where

- the father died without disease at age of death x_{13} before the family recruitment time R,
- the mother developed the disease at age x_{21} and survived to the time of family recruitment at R,
- the first child, the proband, developed the disease at age x_{01} and survived to the family recruitment time R, and
- the second child was disease-free and alive at the family recruitment time R.

The likelihood contribution (5.2.6) of the proband can written as

$$P(\bar{Z}_0(A_0)|Z_0(C_0)=1,C_0,B_0;\phi)=\frac{P(\bar{Z}_0(A_0),B_0;\phi)}{P(Z_0(C_0)=1,C_0,B_0;\phi)}. \tag{5.A.1}$$

Then, numerator of (5.A.1) is given as

$$\begin{aligned}&P(\bar{Z}_0(A_0),B_0;\phi)=P(X_{01}=x_{01},X_{03}>x_{01},X_{02}>A_0,B_0;\phi)\\&=\lambda_1(x_{01})\exp\left(-\int_0^{x_{01}}\lambda_1(s)ds\right)\exp\left(-\int_0^{x_{01}}\lambda_3(B_0+s,s)ds\right)\exp\left(-\int_{x_{01}}^{A_0}\lambda_2(B_0+s,s)ds\right),\end{aligned}$$

and the denominator of (5.A.1) is given as

$$\begin{aligned}&P(Z_0(C_0)=1,C_0,B_0;\phi)=P(X_{01}<C_0,X_{01}<X_{03},X_{02}>C_0,B_0;\phi) \qquad (5.A.2)\\&=\int_0^{C_0}\lambda_1(s)\exp(-\int_0^s\lambda_1(u)du)\exp(-\int_0^s\lambda_3(B_0+u,u)du)\exp(-\int_s^{C_0}\lambda_2(B_0+u,u)du)ds.\end{aligned}$$

In design I, we include information from all family members where we assume that the disease history of the father is available. The contribution of non-probands given the proband data in (5.2.9) is then

$$\prod_{1\le j<l\le 3}\{L^I_{jl}(\psi)\}^{1/2}=\prod_{1\le j<l\le 3}\{P(\bar{\boldsymbol{Z}}^-_{jl}(\boldsymbol{A}^-_{jl})|\bar{Z}_0(A_0),Z_0(A_0)=1,\boldsymbol{A}_{jl},\boldsymbol{B}_{jl};\psi)\}^{1/2} \quad (5.A.3)$$

which is given explicitly as

$$\begin{aligned}&\{P(X_{11}>x_{13},X_{13}=x_{13},X_{21}=x_{21},X_{23}>x_{21},X_{22}>A_2|X_{01}=x_{01},X_{03}>x_{01},X_{02}>A_0;\psi)\\&\times P(X_{11}>x_{13},X_{13}=x_{13},X_{31}>A_3,X_{33}>A_3|X_{01}=x_{01},X_{03}>x_{01},X_{02}>A_0;\psi)\\&\times P(X_{21}=x_{21},X_{23}>x_{21},X_{22}>A_2,X_{31}>A_3,X_{33}>A_3|X_{01}=x_{01},X_{03}>x_{01},X_{02}>A_0;\psi)\}^{1/2}.\end{aligned}$$

As an example, we show how to calculate the first term in above. Using the Clayton copula and its invariance property in (5.2.2), we can write

$$\begin{aligned}&P(X_{11}>x_{13},X_{13}=x_{13},X_{21}=x_{21},X_{23}>x_{21},X_{22}>A_2|X_{01}=x_{01},X_{03}>x_{01},X_{02}>A_0;\psi)\\&=\frac{\partial C(\mathcal{F}(x_{13}|X_{01}=x_{01};\phi_1,\rho),\mathcal{F}(u|X_{01}=x_{01};\phi_1,\rho);\rho^*)}{\partial u}\bigg|_{u=x_{21}}\exp\left(-\int_0^{x_{13}}\lambda_3(B_1+s,s)ds\right)\\&\times\lambda_3(B_1+x_{13},x_{13})\exp\left(-\int_0^{x_{21}}\lambda_3(B_2+s,s)ds\right)\exp\left(-\int_{x_{21}}^{A_2}\lambda_2(B_2+s,s)ds\right),\end{aligned}$$

where $\rho^*=\rho/(1+\rho)$, and $\mathcal{F}(u|X_{01}=x_{01};\phi_1,\rho)=C(\mathcal{F}(u;\phi_1),\mathcal{F}(x_{01};\phi_1);\rho)/\mathcal{F}(x_{01};\phi_1)$.

In design II, we exclude information from the father since he died before the family recruitment time R. The contribution to the augmented composite likelihood can then be written as

$$L^{II}_{23}(\psi)=\frac{P(\bar{\boldsymbol{Z}}^-_{23}(\boldsymbol{A}^-_{23})|\bar{Z}_0(A_0),Z_0(A_0)=1,\boldsymbol{A}_{23},\boldsymbol{B}_{23};\psi)}{P(\boldsymbol{Z}^-_{23}(\boldsymbol{A}^-_{23})\in\{0,1\}^2|\bar{Z}_0(A_0),Z_0(A_0)=1,\boldsymbol{A}_{23},\boldsymbol{B}_{23};\psi)},$$

where the numerator is given as

$$P(X_{21}=x_{21},X_{23}>x_{21},X_{22}>A_2,X_{31}>A_3,X_{33}>A_3|X_{01}=x_{01},X_{03}>x_{01},X_{02}>A_0;\psi),$$

and the denominator is given as

$$
\begin{aligned}
&P(X_{21} > A_2, X_{23} > A_2, X_{31} > A_3, X_{33} > A_3 | X_{01} = x_{01}, X_{03} > x_{01}, X_{02} > A_0; \psi) \\
&+ P(X_{21} < A_2, X_{23} > A_2, X_{22} > A_2, X_{31} > A_3, X_{33} > A_3 | X_{01} = x_{01}, X_{03} > x_{01}, X_{02} > A_0; \psi) \\
&+ P(X_{21} > A_2, X_{23} > A_2, X_{31} < A_3, X_{33} > A_3, X_{32} > A_3 | X_{01} = x_{01}, X_{03} > x_{01}, X_{02} > A_0; \psi) \\
&+ P(X_{21} < A_2, X_{23} > A_2, X_{22} > A_2, X_{31} < A_3, X_{33} > A_3, X_{32} > A_3 | X_{01} = x_{01}, X_{03} > x_{01}, X_{02} > A_0; \psi).
\end{aligned} \tag{5.A.4}
$$

We here show how to obtain the second and the last term in (5.A.4) since the first term in (5.A.4) is straightforward.

$$
\begin{aligned}
&P(X_{21} < A_2, X_{23} > A_2, X_{22} > A_2, X_{31} > A_3, X_{33} > A_3 | X_{01} = x_{01}, X_{03} > x_{01}, X_{02} > A_0; \psi) \\
&= \int_0^{A_2} \frac{\partial C(\mathcal{F}(u|X_{01} = x_{01}; \phi_1, \rho), \mathcal{F}(A_3|X_{01} = x_{01}; \phi_1, \rho); \rho^*)}{\partial u}\bigg|_{u=s} \exp\left(-\int_0^s \lambda_3(B_2 + v, v)dv\right) \\
&\times \exp\left(-\int_s^{A_2} \lambda_2(B_2 + v, v)dv\right) \exp\left(-\int_0^{A_3} \lambda_3(B_3 + v, v)dv\right) ds,
\end{aligned}
$$

and

$$
\begin{aligned}
&P(X_{21} < A_2, X_{23} > A_2, X_{22} > A_2, X_{31} < A_3, X_{33} > A_3, X_{32} > A_3 | X_{01} = x_{01}, X_{03} > x_{01}, X_{02} > A_0; \psi) \\
&= \int_0^{A_3} \int_0^{A_2} \frac{\partial^2 C(\mathcal{F}(u|X_{01} = x_{01}; \phi_1, \rho), \mathcal{F}(w|X_{01} = x_{01}; \phi_1, \rho); \rho^*)}{\partial u \partial w}\bigg|_{u=s, w=y} \\
&\times \exp\left(-\int_0^s \lambda_3(B_2 + v, v)dv\right) \exp\left(-\int_s^{A_2} \lambda_2(B_2 + v, v)dv\right) \exp\left(-\int_0^y \lambda_3(B_2 + v, v)dv\right) \\
&\times \exp\left(-\int_y^{A_3} \lambda_2(B_3 + v, v)dv\right) dsdy.
\end{aligned}
$$

We suppose that two auxiliary data are used as introduced in Section 5.3. We consider a particular individual in registry data who developed the disease at age x_{r1} and died at age x_{r2}. The likelihood contribution is

$$
P(\bar{Z}_r(A_r^*)|Z_r(C_r) = 1, C_r, B_r; \phi) = \frac{P(\bar{Z}_r(x_{r3}), B_0; \phi)}{P(Z_r(C_r) = 1, C_r, B_r; \phi)}. \tag{5.A.5}
$$

Then, numerator of (5.A.5) is given as

$$P(\bar{Z}_0(x_{r3}), B_r; \phi) = P(X_{r1} = x_{r1}, X_{r3} > x_{r1}, X_{r2} = x_{r2}, B_r; \phi)$$
$$= \lambda_1(x_{r1}) \exp\left(-\int_0^{x_{r1}} \lambda_1(s)ds\right) \exp\left(-\int_0^{x_{r1}} \lambda_3(B_r + s, s)ds\right) \exp\left(-\int_{x_{r1}}^{x_{r2}} \lambda_2(B_r + s, s)ds\right),$$

and the denominator of (5.A.5) has the same form as (5.A.2) From the cross-sectional survey, we consider an individual who developed the disease by the age at contact for survey C_r. The likelihood contribution is

$$P(Z_r(C_r) = 1 | Z_r(C_r) \in \{0, 1\}, B_r) = \frac{P(Z_r(C_r) = 1, B_r; \phi)}{P(Z_r(C_r) = 0; \phi) + P(Z_r(C_r) = 1; \phi)}$$

where

$$P(Z_r(C_r) = 0, B_r; \phi) = \exp\left(-\int_0^{C_r} \lambda_1(u)du\right) \exp\left(-\int_0^{C_r} \lambda_3(B_0 + u, u)du\right),$$

and

$$P(Z_r(C_r) = 1, B_r; \phi) = \int_0^{C_r} \lambda_1(s) \exp\left(-\int_0^{s} \lambda_1(u)du\right) \exp\left(-\int_0^{s} \lambda_3(B_0 + u, u)du\right)$$
$$\times \exp\left(-\int_s^{C_r} \lambda_2(B_0 + u, u)du\right) ds.$$

To compute a one dimensional integral we use the `intergral` function in `R`, and for two dimensional integrals we directly code Gaussian-Quadrature a numerical integration algorithm with 40 nodes.

Appendix 5.B. Calculation of $P(G_{ijl})$

Recall $G_{ijl} = (G_{i0}, G_{ij}, G_{il})'$ is a vector of genetic markers for the proband and members j and l of family i; for families with two members we let $G_{ij} = (G_{i0}, G_{ij})'$. We can calculate $P(\boldsymbol{G}_{ijl})$ based on the assumption that the process is in Hardy-Weinberg equilibrium and following Mendel's law, with a risk allele frequency p (Elandt-Johnson, 1971). Here, we

consider two or three members per family for the pairwise likelihood in which we denote G_p, G_c the genotype of a parent as G_p and the genotype of the child as G_s.

Joint distribution of alleles for different types of pairs of family members

$\boldsymbol{G}$		$P(G_p, G_p)$	$P(G_p, G_c)$	$P(G_c, G_c)$
1	1	$(1-q^2)^2$	p^2q+p	$\frac{1}{4}p^2(1+p)^2+pq(2p+1)$
1	0	$(1-q^2)q^2$	pq^2	$\frac{1}{4}p^2q^2+\frac{1}{2}pq^2(1+q)$
0	1	$(1-q^2)q^2$	pq^2	$\frac{1}{4}p^2q^2+\frac{1}{2}pq^2(1+q)$
0	0	q^4	q^3	$\frac{1}{4}q^2(1+q)^2$

Joint distribution of alleles for different types of triples of family members

$\boldsymbol{G}$			$P(G_{p_1}, G_{p_2}, G_c)$	$P(G_p, G_{c_1}, G_{c_2})$	$P(G_{c_1}, G_{c_2}, G_{c_3})$
1	1	1	$p^2(1+2q)$	$\frac{1}{4}p^2(1+p)(5-3p)+\frac{1}{2}pq(p+pq+1)$	$\frac{1}{16}p^2(1+3p)(7-3p)+\frac{1}{4}pq(6p+3pq+2)$
1	1	0	p^2q^2	$\frac{1}{4}p^2q^2+\frac{1}{2}pq^2$	$\frac{5}{16}p^2q^2+\frac{1}{4}pq^2(1+q)$
1	0	1	pq^2	$\frac{1}{4}p^2q^2+\frac{1}{2}pq^2$	$\frac{5}{16}p^2q^2+\frac{1}{4}pq^2(1+q)$
1	0	0	pq^3	$\frac{1}{4}pq^2(1+q)$	$\frac{1}{16}p^2q^2+\frac{1}{8}pq^2(1+3q)$
0	1	1	pq^2	$\frac{1}{2}pq^2(1+p)$	$\frac{5}{16}p^2q^2+\frac{1}{4}pq^2(1+q)$
0	1	0	pq^3	$\frac{1}{2}pq^3$	$\frac{1}{16}p^2q^2+\frac{1}{8}pq^2(1+3q)$
0	0	1	0	$\frac{1}{2}pq^3$	$\frac{1}{16}p^2q^2+\frac{1}{8}pq^2(1+3q)$
0	0	0	q^4	$\frac{1}{2}q^3(1+q)$	$\frac{1}{16}q^2(1+3q)^2$

Table 5.6: Joint probability model for genetic markers for two (top) or three (bottom) family members according to their relationships

Chapter 6

Remarks and Future Research

6.1 Overview

This thesis develops new statistical methods for different types of data arising from life history processes to address heterogeneity and dependence. In Chapter 2, motivated by the common setting in which recurrent exacerbations arise of an appreciable duration, we propose a bivariate random effects model for an alternating two-state process. The time scale governing onset of exacerbations is the time since the onset of the process (*i.e.* it is conditionally Markov) whereas a conditionally semi-Markov intensity is used for recovery from exacerbations. The individual-specific random effects address both heterogeneity for each type of transition, and a copula model is used to accommodate dependence between the two component random effects. Different combinations of marginal random effect distributions and copula functions are considered, and the effect of misspecification of the copula function is examined.

In Chapter 3, we consider the setting of a randomized clinical trial where the onset and resolution of recurrent exacerbations are governed by the two-state process of Chapter 2, but where semiparametric marginal rate-based models are used for the analysis of the recurrent onset times only. This is motivated by the fact that such data are routinely analysed in this way in clinical trials but in such settings the recurrent event analysis means

that misspecified models are being used. We examine the effect of using two different risk set definitions, including or excluding individuals from the risk set during symptomatic periods. We also study the impact of misspecification of risk sets on power in a clinical trial since this represents the area of most common application.

The analysis of multiple types of recurrent events arising due to the same cause was addressed in Chapter 4. A multivariate mixed-Poisson model was adopted to accommodate between-individual variation in the event and copula functions were used to link event type-specific random effects to capture dependence between the different types of events. A semiparametric estimation procedure is developed via an expectation-maximization algorithm. For more than 2 types of events, inferences are described based on pairwise composite likelihood with both simultaneous and two-stage estimation procedures.

In Chapter 5, attention was directed to modeling familial aggregation in the (possibly latent) age of onset in framework of a marginal illness-death model. The dependence structure for the (latent) age of disease onset between family members is modeled again via copula functions. In family studies, biased sampling schemes are typically employed in the recruitment of family members. An individual with disease is recruited first to the study where the right-truncated disease onset times and the left-truncated survival times are observed. Then, families of this individual, called the proband, are recruited. Depending on the study design, complete retrospective data of non-probands may be available, or only data from family members who can attend a clinic may be available. Both designs involve biased sampling and raise identifiability and estimability issues. Auxiliary data is of use to address this issue and improve efficiency of estimators. To study the genetic basis of disease, we also accommodate genetic factors as covariates while missing genotypes were addressed.

In the following sections we outline further research for each topic.

6.2 Ongoing Work in Alternating Two-state Processes

Several extensions are possible for alternating two-state processes. When the mortality comes into play as is the case with chronic obstructive pulmonary disease in elderly indi-

viduals for example,a death state may be added as an absorbing state to create a three state process. The cumulative mean function should condition on survival states explicitly or marginalized over the possible times of death in this context. In the application in Chapter 2, sampled individuals are required to have experienced hospitalizations and individual data are accessible from the beginning of hospitalization over the subsequent study period. The model of Chapter 2 can be generalized to reflect this sampling scheme in the likelihood construction.

In the motivating example in Chapter 3, individuals in treatment and control arms received the same treatment during symptomatic periods. Therefore, to examine the efficacy of treatment, a focus is the treatment effect on the onset of exacerbation. However it is challenging to obtain robust multiplicative rate-based treatment effect estimators on symptom onset because the models are typically misspecified in a way that inconsistent estimates are obtained. Utility-based model (Cook *and others*, 2003) or methods based on the expected length of time in a particular state (Grand and Putter, 2016) are potential frameworks worthy of further consideration.

6.3 Ongoing Work in Multi-type Recurrent Events

This general approach can be naturally extended to accommodate multi-type interval-censored recurrent event data of the sort studied by Chen *and others* (2005) where the exact event times are unavailable but counts of the number of events in consecutive intervals of each type is known. The fact that this method is implemented using an expectation-maximization algorithm means that it can also be naturally generalized to accommodate settings where the event types are partially missing; see Chen and Cook (2009). An alternative framework for analyzing multi-type recurrent events is via marginal methods and estimating functions (Cai and Schaubel, 2004). While this can be appealing because it is based on partially specified models it does not lend itself naturally to prediction. Fully specified models, even when fitted under composite likelihood, can facilitate prediction of future events of any type or all types, and exploits the history of the joint processes which should yield more accurate and precise predictions of features of interest (Fredette and

Lawless, 2007).

Another extension is to obtain a global estimate of a treatment effect on multi-type events. Several approaches summarized in Wei and Glidden (1997) can be exploited for global test statistics and we can develop sample size calculation for multiple events. This was not explored here.

Recently, Claggett *and others* (2018) proposed a method for nonparametric inference for multiple events based on a reverse counting process. This approach is appealing to have a simple interpretation of a global treatment effect on individual's disease process. This method could be adopted to deal with recurrent events where the events occur repeatedly.

6.4 Ongoing Work in Family Studies

There are several exciting extensions for the works on family studies. A natural extension is to consider multivariate genetic markers which may affect the age of disease onset. This leads to computational burden due to the summation of all possible combination of genetic markers for missing genotypes. We may use other sources of population studies to calculate the allele frequency and we may exploit this value to assume that the allele frequency p is known in our proposed model. This will reduce the number of parameters to estimate.

Another extension is to allow a different dependence structure in families using, for example, the Gaussian copula function; this was done in Zhong and Cook (2016). The degree of dependence may differ depending on which pair of individuals is being considered; a weaker dependence is expected between parents than parent-child pairs from a genetic point of view. However, in that case, a simple form of a cross ratio is not available.

Two-stage estimation is a plausible extension where registry data, the current status survey data, and mortality rates are used to estimate marginal parameters, and at the second stage a dependence parameter is estimated given the marginal parameter estimates. This approach is natural in that the two sources of auxiliary data do not have information with respect to the familial association.

Finally one may also examine how much power improves when testing genetic effects by using additional current status survey data in design II. As shown in simulation studies,

the efficiency of estimators increases with the survey data in design II even if sampled individuals are not genotyped.

References

Aalen, O., Borgan, O. and Gjessing, H. (2008). *Survival and Event History Analysis: a Process Point of View*, Statistics for biology and health. New York, NY: Springer.

Aalen, O. O. (2012). Armitage lecture 2010: understanding treatment effects: the value of integrating longitudinal data and survival analysis. *Statistics in Medicine* **31**(18), 1903–1917.

Aalen, O. O., Borgan, Ø., Keiding, N. and Thormann, J. (1980). Interaction between life history events. nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics* **7**(4), 161–171.

Aalen, O. O., Cook, R. J. and Røysland, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis* **21**(4), 579–593.

Ambrus, J. L. (2004). Nutrition and infectious diseases in developing countries and problems of acquired immunodeficiency syndrome. *Experimental Biology and Medicine* **229**(6), 464–472.

Andersen, E. W. (2004). Composite likelihood and two-stage estimation in family studies. *Biostatistics* **5**(1), 15–30.

Andersen, P. K. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in Medicine* **7**(6), 661–670.

Andersen, P. K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N. and Kreiner, S. (1985). A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics* **41**(4), 921–932.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10**, 1100–1120.

Anderson, V. E. (1961). Statistical studies of probands and their relatives. *Annals of the New York Academy of Sciences* **91**(1), 781–796.

Balliu, B., Tsonaka, R., van der Woude, D., Boehringer, S. and Houwing-Duistermaat, J. J. (2012). Combining family and twin data in association studies to estimate the noninherited maternal antigens effect. *Genetic Epidemiology* **36**(8), 811–819.

Bandeen-Roche, K. and Liang, K. (2002). Modelling multivariate failure time associations in the presence of a competing risk. *Biometrika* **89**(2), 299–314.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B* **36**(2), 192–236.

Beyersmann, J., Dettenkofer, M., Bertz, H. and Schumacher, M. (2007). A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards. *Statistics in Medicine* **26**(30), 5360–5369.

Borer, J. S., Böhm, M., Ford, I., Komajda, M., Tavazzi, L., Sendon, J. L., Alings, M., Lopez-de Sa, E., Swedberg, K. and Investigators, Shift. (2012). Effect of ivabradine on recurrent hospitalization for worsening heart failure in patients with chronic systolic heart failure: the shift study. *European Heart Journal* **33**(22), 2813–2820.

Cai, J. and Schaubel, D. E. (2004). Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Analysis* **10**(2), 121–138.

Chandran, V. (2013). The genetics of psoriasis and psoriatic arthritis. *Clinical Reviews in Allergy & Immunology* **44**(2), 149–156.

Chandran, V. and Raychaudhuri, S. P. (2010). Geoepidemiology and environmental factors of psoriasis and psoriatic arthritis. *Journal of Autoimmunity* **34**(3), J314–J321.

Chandran, V., Schentag, C. T., Brockbank, J. E., Pellett, F. J., Shanmugarajah, S., Toloza, S. MA., Rahman, P. and Gladman, D. D. (2009). Familial aggregation of psoriatic arthritis. *Annals of the Rheumatic Diseases* **68**(5), 664–667.

Chatterjee, N., Kalaylioglu, Z., Shih, J. H. and Gail, M. H. (2006). Case–control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics* **62**(1), 36–48.

Chen, B. E. and Cook, R. J. (2009). The analysis of multivariate recurrent events with partially missing event types. *Lifetime Data Analysis* **15**(1), 41.

Chen, B. E., Cook, R. J., Lawless, J. F. and Zhan, M. (2005). Statistical methods for multivariate interval-censored recurrent events. *Statistics in Medicine* **24**(5), 671–691.

Chen, X., Wang, Q., Cai, J. and Shankar, V. (2012). Semiparametric additive marginal regression models for multiple type recurrent events. *Lifetime Data Analysis* **18**(4), 504–527.

Cheng, Y., Fine, J. P. and Kosorok, M. R. (2009). Nonparametric association analysis of exchangeable clustered competing risks data. *Biometrics* **65**(2), 385–393.

Claggett, B., Tian, L., Fu, H., Solomon, S. D. and Wei, L. J. (2018). Quantifying the totality of treatment effect with multiple event-time observations in the presence of a terminal event from a comparative clinical study. *Statistics in Medicine* **37**(25), 3589–3598.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**(1), 141–151.

Cole, S. R, Platt, R. W, Schisterman, E. F, Chu, H., Westreich, D., Richardson, D. and Poole, C. (2009). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* **39**(2), 417–420.

Cook, R. J. and Lawless, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16**(8), 911–924.

Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. New York, NY: Springer.

Cook, R. J. and Lawless, J. F. (2013). Concepts and tests for trend in recurrent event processes. *Journal of The Iranian Statistical Society* **12**(1), 35–69.

Cook, R. J. and Lawless, J. F. (2014). Statistical issues in modeling chronic disease in cohort studies. *Statistics in Biosciences* **6**(1), 127–161.

Cook, R. J. and Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. New York: Chapman and Hall/CRC.

Cook, R. J., Lawless, J. F., Lakhal-Chaieb, L. and Lee, K. (2009). Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone. *Journal of the American Statistical Association* **104**(485), 60–75.

Cook, R. J., Lawless, J. F. and Lee, K. (2003). Cumulative processes related to event histories. *SORT-Statistics and Operations Research Transactions* **27**(1), 13–30.

Cook, R. J., Lawless, J. F. and Lee, K. (2010). A copula-based mixed poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine* **29**(6), 694–707.

Cook, R. J., Ng, E., Mukherjee, J. and Vaughan, D. (1999). Two-state mixed renewal processes for chronic disease. *Statistics in Medicine* **18**(2), 175–188.

Cortese, G. and Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal* **52**(1), 138–158.

Cox, D. R. (1967). *Renewal Theory*. London: Methuen.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. London: Chapman and Hall.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *The Annals of Statistics* **16**(4), 1475–1489.

Datta, S., Satten, G. A. and Datta, S. (2000). Nonparametric estimation for the three-stage irreversible illness–death model. *Biometrics* **56**(3), 841–847.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B* **39**(1), 1–38.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. New York, NY: Springer.

Eder, L., Chandran, V., Pellett, F., Shanmugarajah, S., Rosen, C. F., Bull, S. B. and Gladman, D. D. (2012). Differential human leucocyte allele association between psoriasis and psoriatic arthritis: a family-based association study. *Annals of the Rheumatic Diseases* **71**(8), 1361–1365.

Elandt-Johnson, R. C. (1971). Joint genotype distributions of s children and a parent, and of s siblings: multiple alleles. *American Journal of Human Genetics* **23**(5), 442.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**(446), 496–509.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* **6**(1), 13–25.

Fredette, M. and Lawless, J. F. (2007). Finite-horizon prediction of recurrent events, with application to forecasts of warranty claims. *Technometrics* **49**(1), 66–80.

Gao, X. and Song, P. K. (2011). Composite likelihood em algorithm with applications to multivariate hidden markov model. *Statistica Sinica* **21**(1), 165–185.

Garber, J., Kriss, M. R., Koch, M. and Lindholm, L. (1988). Recurrent depression in adolescents: A follow-up study. *Journal of the American Academy of Child & Adolescent Psychiatry* **27**(1), 49–54.

Gelfand, J. M., Gladman, D. D., Mease, P. J., Smith, N., Margolis, D. J., Nijsten, T., Stern, R. S., Feldman, S. R. and Rolstad, T. (2005*a*). Epidemiology of psoriatic arthritis in the population of the United States. *Journal of the American Academy of Dermatology* **53**(4), 573–e1.

Gelfand, J. M., Weinstein, R., Porter, S. B., Neimann, A. L., Berlin, J. A. and Margolis, D. J. (2005*b*). Prevalence and treatment of psoriasis in the United Kingdom: a population-based study. *Archives of Dermatology* **141**(12), 1537–1541.

Genest, C. and Rivest, L. P. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association* **88**(423), 1034–1043.

Ghosh, D. and Lin, D. Y. (2000). Nonparametric analysis of recurrent events and death. *Biometrics* **56**(2), 554–562.

Ghosh, D. and Lin, D. Y. (2002). Marginal regression models for recurrent and terminal events. *Statistica Sinica* **12**(3), 663–688.

Gladman, D. D. (1991). Psoriatic arthritis. In: *Prognosis in the Rheumatic Diseases*. Dordrecht: Springer, pp. 153–166.

Gladman, D. D. (2008). Mortality in psoriatic arthritis. *Clinical & Experimental Rheumatology* **26**(5), S62.

Gladman, D. D. and Farewell, V. T. (1995). The role of HLA antigens as indicators of disease progression in psoriatic arthritis. *Arthritis and Rheumatology* **38**(6), 845–850.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* **31**(4), 1208–1211.

Gong, G., Hannon, N. and Whittemore, A. S. (2010). Estimating gene penetrance from family data. *Genetic Epidemiology* **34**(4), 373–381.

Grand, M. K. and Putter, H. (2016). Regression models for expected length of stay. *Statistics in Medicine* **35**(7), 1178–1192.

Grossman, R., Mukherjee, J., Vaughan, D., Eastwood, C., Cook, R. J., LaForge, J. and Lampron, N. (1998). A 1-year community-based health economic study of ciprofloxacin vs usual antibiotic treatment in acute exacerbations of chronic bronchitis: the Canadian Ciprofloxacin Health Economic Study Group. *Chest* **113**(1), 131–141.

He, W. and Lawless, J. F. (2003). Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics* **59**(4), 837–848.

Hernán, M. A. and Robins, J. (2016). *Causal Inference*. Boca Raton: Chapman & Hall/CRC.

Hortobagyi, G. N. (1998). Treatment of breast cancer. *New England Journal of Medicine* **339**(14), 974–984.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Analysis* **5**(3), 239–264.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York, NY: Springer New York.

Hougaard, P., Harvald, B., Holm, N. V., Flournoy, N., Islam, M. A and Singh, K. P. (1992). Assessment of dependence in the life times of twins. In: *Survival Analysis: State of the Art*. Dordrecht: Springer, pp. 77–97.

Hsu, L., Chen, L., Gorfine, M. and Malone, K. (2004). Semiparametric estimation of marginal hazard function from case–control family studies. *Biometrics* **60**(4), 936–944.

Hsu, L. and Gorfine, M. (2005). Multivariate survival analysis for case–control family data. *Biostatistics* **7**(3), 387–398.

Hu, X. J., Lorenzi, M., Spinelli, J. J., Ying, S. C. and McBride, M. L. (2011). Analysis of recurrent events with non-negligible event duration, with application to assessing hospital utilization. *Lifetime Data Analysis* **17**(2), 215–233.

Jayaram, L., Pizzichini, M. M., Cook, R. J., Boulet, L. P., Lemiere, C., Pizzichini, E., Cartier, A., Hussack, P., Goldsmith, C. H., Laviolette, M. *and others*. (2006). Determining asthma treatment by monitoring sputum cell counts: effect on exacerbations. *European Respiratory Journal* **27**(3), 483–494.

Jiang, F. and Haneuse, S. (2017). A semi-parametric transformation frailty model for semi-competing risks survival data. *Scandinavian Journal of Statistics* **44**(1), 112–129.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Boca Raton, FL: CRC Press.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*, Volume 360. Hoboken, NJ, USA: John Wiley & Sons.

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Phil. Trans. R. Soc. Lond. A* **332**(1627), 487–509.

Keiding, N. (2006). Event history analysis and the cross-section. *Statistics in Medicine* **25**(14), 2343–2364.

Kessing, L. V., Hansen, M. G. and Andersen, P. K. (2004). Course of illness in depressive and bipolar disorders. *The British Journal of Psychiatry* **185**(5), 372–377.

Kessing, L. V., Olsen, E. W., Andersen, P. K. and in cooperation with the Department of Psychiatric Demography, Psychiatric Hospital Risskov Denmark University of Aarhus. (1999). Recurrence in affective disorder: analyses with frailty models. *American Journal of Epidemiology* **149**(5), 404–411.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **43**(3), 795–806.

Lakhal-Chaieb, L., Cook, R. J. and Zhong, Y. (2018). Testing the heritability and parent-of-origin hypotheses for ages at onset of psoriatic arthritis under biased sampling. *Biometrics under revision*.

Langley, R. G. B, Krueger, G. G. and Griffiths, C. E. M. (2005). Psoriasis: epidemiology, clinical features, and quality of life. *Annals of the Rheumatic Diseases* **64**(suppl 2), ii18–ii23.

Latouche, A., Boisson, Vé., Chevret, S. and Porcher, R. (2007). Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine* **26**(5), 965–974.

Lawless, J. F. (1987*a*). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15**(3), 209–225.

Lawless, J. F. (1987*b*). Regression methods for Poisson process data. *Journal of the American Statistical Association* **82**(399), 808–815.

Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**(2), 158–168.

Leclerc, M., EMBRACE Investigators, GEMO Study Collaborators, INHERIT Investigators, Antoniou, A. C., Simard, J. and Lakhal-Chaieb, L. (2015). Analysis of multivariate failure times in the presence of selection bias with application to breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**(3), 525–541.

Lee, E. W., Wei, L. J., Amato, D. A. and Leurgans, S. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: *Survival Analysis: State of the Art*. Dordrecht: Springer, pp. 237–247.

Lee, J. and Cook, R. J. (2018). Heterogeneity and dependence modeling for alternating two-state processes via copulas. *Manuscript*.

Lemaire, M., Islam, Q. S., Shen, H., Khan, M. A., Parveen, M., Abedin, F., Haseen, F., Hyder, Z., Cook, R. J. and Zlotkin, S. H. (2011). Iron-containing micronutrient powder provided to children with moderate-to-severe malnutrition increases hemoglobin concentrations but not the risk of infectious morbidity: a randomized, double-blind, placebo-controlled, noninferiority safety trial. *The American Journal of Clinical Nutrition* **94**(2), 585–593.

Li, H., Yang, P. and Schwartz, A. G. (1998). Analysis of age of onset data from case-control family studies. *Biometrics* **54**(3), 1030–1039.

Liang, K. and Beaty, T. H. (2000). Statistical designs for familial aggregation. *Statistical Methods in Medical Research* **9**(6), 543–562.

Liang, K., Self, S. G. and Chang, Y. (1993). Modelling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society. Series B* **55**(2), 441–453.

Lin, D. Y., Sun, W. and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika* **86**(1), 59–70.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B* **62**(4), 711–730.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**(1), 221–39.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B* **44**(2), 226–233.

Mazroui, Y., Mathoulin-Pélissier, S., MacGrogan, G., Brouste, Vé. and Rondeau, V. (2013). Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data. *Biometrical Journal* **55**(6), 866–884.

McCulloch, C. E. and Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics* **67**(1), 270–279.

Mesfioui, M. and Quessy, J. (2008). Dependence structure of conditional Archimedean copulas. *Journal of Multivariate Analysis* **99**(3), 372–385.

Moll, J. M. and Wright, V. (1973). Familial occurrence of psoriatic arthritis. *Annals of the Rheumatic Diseases* **32**(3), 181.

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer.

Nelson, W. (1995). Confidence limits for recurrence data-applied to cost or number of product repairs. *Technometrics* **37**(2), 147–157.

Ng, E. and Cook, R. J. (1997). Modeling two-state disease processes with random effects. *Lifetime Data Analysis* **3**(4), 315–335.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**(406), 487–493.

O'Keeffe, A. G., Tom, B. DM. and Farewell, V. T. (2011). A case-study in the clinical epidemiology of psoriatic arthritis: multistate models and causal arguments. *Journal of the Royal Statistical Society: Series C* **60**(5), 675–699.

Olesen, A. V. and Parner, E. T. (2006). Correcting for selection using frailty models. *Statistics in Medicine* **25**(10), 1672–1684.

Parner, E. *and others*. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* **26**(1), 183–214.

Pedersen, O. B., Svendsen, A. Jø., Ejstrup, L., Skytthe, A. and Junker, P. (2008). On the heritability of psoriatic arthritis. Disease concordance among monozygotic and dizygotic twins. *Annals of the Rheumatic Diseases* **67**(10), 1417–1421.

Pfeiffer, R. M., Pee, D. and Landi, M. T. (2008). On combining family and case-control studies. *Genetic Epidemiology* **32**(7), 638–646.

Pollock, R., Chandran, V., Barrett, J., Eder, L., Pellett, F., Yao, C., Lino, M., Shanmugarajah, S., Farewell, V. T. and Gladman, D. D. (2011). Differential major histocompatibility complex class i chain-related a allele associations with skin and joint manifestations of psoriatic disease. *Tissue Antigens* **77**(6), 554–561.

Pollock, R. A., Thavaneswaran, A., Pellett, F., Chandran, V., Petronis, A., Rahman, P. and Gladman, D. D. (2015). Further evidence supporting a parent-of-origin effect in psoriatic disease. *Arthritis Care & Research* **67**(11), 1586–1590.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**(4), 541–554.

Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**(2), 373–379.

Putter, H., Fiocco, M. and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* **26**(11), 2389–2430.

Putter, H. and van Houwelingen, H. C. (2015). Frailties in multi-state models: Are they identifiable? do we need them? *Statistical Methods in Medical Research* **24**(6), 675–692.

Queiro, R., Morante, I., Cabezas, I. and Acasuso, B. (2015). HLA-B27 and psoriatic disease: a modern view of an old relationship. *Rheumatology* **55**(2), 221–229.

Queiro, R., Sarasqueta, C., Torre, J., Tinturé, T. and Lopez-Lagunas, I. (2001). Comparative analysis of psoriatic spondyloarthropathy between men and women. *Rheumatology International* **21**(2), 66–68.

Reveille, J. D., Hirsch, R., Dillon, C. F., Carroll, M. D. and Weisman, M. H. (2012). The prevalence of HLA–B27 in the US: data from the US national health and nutrition examination survey, 2009. *Arthritis & Rheumatology* **64**(5), 1407–1411.

Robert, B. (2017). Mortality data for Canada. https://www.mortality.org/cgi-bin/hmd/country.php?cntr=CAN&level=1.

Romanowski, B., Marina, R. B., Roberts, J. N., Valtrex HS230017 Study Group *and others*. (2003). Patients' preference of valacyclovir once-daily suppressive therapy versus twice-daily episodic therapy for recurrent genital herpes: a randomized study. *Sexually Transmitted Diseases* **30**(3), 226–231.

Schaible, U. E. and Stefan, H. E. (2007). Malnutrition and infection: complex mechanisms and global impacts. *PLoS Med* **4**(5), e115.

Scheike, T. H., Holst, K. K. and Hjelmborg, J. B. (2014). Estimating heritability for cause specific mortality based on twin studies. *Lifetime Data Analysis* **20**(2), 210–233.

Scheike, T. H. and Sun, Y. (2012). On cross-odds ratio for multivariate competing risks data. *Biostatistics* **13**(4), 680–694.

Scheike, T. H., Sun, Y., Zhang, M. and Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks data. *Biometrika* **97**(1), 133–145.

Scheike, T. H., Zhang, M. J. and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95**(1), 205–220.

Schweder, T. (1970). Composable Markov processes. *Journal of Applied Probability* **7**(2), 400–410.

Shih, J. H. and Albert, P. S. (2010). Modeling familial association of ages at onset of disease in the presence of competing risk. *Biometrics* **66**(4), 1012–1023.

Shih, J. H. and Chatterjee, N. (2002). Analysis of survival data from case–control family studies. *Biometrics* **58**(3), 502–509.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51**(4), 1384–1399.

The Diabetes Control and Complications Trial Research Group. (1986). The Diabetes Control and Complications Trial (DCCT): design and methodologic considerations for the feasibility phase. *Diabetes* **35**(5), 530–545.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis* **92**(1), 1–28.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**(1), 5–42.

Wei, L., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**(408), 1065–1073.

Wei, L. J. and Glidden, D. V. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine* **16**(8), 833–839.

Wienke, A. (2010). *Frailty Models in Survival Analysis*. Boca Raton, FL: CRC Press.

Wong, K., Gladman, D. D., Husted, J., Long, J. A. and Farewell, V. T. (1997). Mortality studies in psoriatic arthritis. Results from a single outpatient clinic. I. Causes and risk of death. *Arthritis & Rheumatology* **40**(10), 1868–1872.

Xu, J., Kalbfleisch, J. D. and Tai, B. (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics* **66**(3), 716–725.

Xue, X. and Brookmeyer, R. (1996). Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Analysis* **2**(3), 277–289.

Yusuf, S., Wittes, J., Probstfield, J. and Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Jama* **266**(1), 93–98.

Zhang, H., Olschwang, S. and Yu, K. (2010). Statistical inference on the penetrances of rare genetic mutations based on a case–family design. *Biostatistics* **11**(3), 519–532.

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics* **33**(3), 335–356.

Zheng, Y., Heagerty, P. J., Hsu, L. and Newcomb, P. A. (2010). On combining family-based and population-based case–control data in association studies. *Biometrics* **66**(4), 1024–1033.

Zhong, Y. and Cook, R. J. (2016). Augmented composite likelihood for copula modeling in family studies under biased sampling. *Biostatistics* **17**(3), 437–452.

Zhong, Y. and Cook, R. J. (2017). Second-order estimating equations for clustered current status data from family studies using response-dependent sampling. *Statistics in Biosciences* **10**(1), 1–24.

Zhou, B., Fine, J., Latouche, A. and Labopin, M. (2012). Competing risks regression for clustered data. *Biostatistics* **13**(3), 371–383.